Proceedings of the Eighteenth Australasian International Conference on Speech Science and Technology

13–16 December 2022 ▪ Canberra, Australia

Conference Information

Table of Contents

Author Index

Search

Editor: Rosey Billington

ASSTA
The Australasian Speech Science & Technology Association Incorporated

Proceedings of the Eighteenth Australasian International Conference on Speech Science and Technology. ISSN 2207-1296. For technical support please contact Causal Productions (info@causalproductions.com).

Welcome to Delegates

On behalf of the organising committee, and the Australasian Speech Science and Technology Association (ASSTA), we welcome you to the 18th Australasian International Conference on Speech Science and Technology (SST2022). We welcome attendees gathering in-person in Canberra, and in addition, attendees joining online, as we hold the first hybrid SST conference. We acknowledge the Ngunnawal and Ngambri people, the traditional owners of the land SST is being hosted on, and pay our respects to their elders past and present.

It has been four years since the SST community last had the opportunity to meet. The onset of the COVID-19 pandemic prevented a 2020 event from taking place, and has presented enormous challenges for many in our community, both personal and professional. Many research projects have been interrupted, and many people have had to pivot to new approaches. SST2022 offers the Australasian speech science and technology community an opportunity to reconnect, share emerging findings, reflect on developments in the field (including new directions taken in response to challenges posed by the pandemic), and welcome new students and colleagues attending SST for the first time.

For SST2022, we invited submission of 4-page papers (for a 20-minute oral presentation or a poster presentation and publication in the proceedings) and 1-page abstracts (for poster presentation only and publication on the conference website). Thanks to the efforts of our qualified reviewers, the submissions were all blind peer reviewed by three anonymous impartial and independent experts. Papers were evaluated on the basis of their relevance to the scientific areas of SST, the clarity of the research goals, approaches and findings, and the novelty and value of the contribution relative to previous research on the topic. Submissions were selected on the basis of reviewer comments and scores. Authors of accepted submissions revised their papers and abstracts on the basis of reviewer comments, and resubmitted final deanonymised versions of their submissions.

The SST conference series has had “International” in its title since the second SST in 1998, reflecting its keynotes from around the world at the cutting edge of research in speech science and technology, and delegates from Australia, New Zealand, and overseas. This year is no exception. Many of the delegates in 2022 are from overseas, hailing from China, Hong Kong, Singapore, Japan, Thailand, New Caledonia, France, Switzerland, the United Kingdom, Canada, the United States and Botswana. We are delighted to welcome our international and local presenters, including keynote speakers Paul Foulkes, Catherine Watson and Phil Rose. We also welcome Paul and Catherine as tutorial day presenters, along with Helen Fraser, Debbie Loakes, and Márton Sóskuthy. We extend a welcome to Márton and to Ghada Khattab as keynote speakers at the SocioPhonAus3 satellite event taking place at SST. Thank you to all guest presenters for contributing to the continuing success of our conference series.

We would like to acknowledge the Australian National University for its support of the conference, in particular the School of Literature, Languages and Linguistics and the School of Culture, History and Language, who provided both logistical and financial support. We

also gratefully acknowledge the financial support of the Research School of Humanities and the Arts, the College of Asia and the Pacific, and the ARC Centre of Excellence for the Dynamics of Language.

The committee members would like to thank all volunteer reviewers who contributed their efforts to the review process and made it possible to develop such a quality technical program. Finally, we thank all authors and attendees for contributing your latest research results and your enthusiasm to share and discuss them with the wider speech science and technology community.

In keeping with SST tradition, this conference has papers from a range of topics across speech science and technology, and allows many opportunities for multidisciplinary engagement, including through two special sessions. We hope that you will find this programme interesting, engaging, and a stimulus to your next research advances, and that you enjoy this long-awaited event.

Rosey Billington

Conference Chair

SST2022 Sponsors

Australian National University

Hosted and organised by:

The School of Literature, Languages and Linguistics

The School of Culture, History and Language

Supported by:

The Research School of Humanities and the Arts

The College of Asia and the Pacific

The ARC Centre of Excellence for the Dynamics of Language

ARC CENTRE OF EXCELLENCE FOR
THE DYNAMICS OF LANGUAGE

SST2022 Conference Organisation

Organising Committee

Rosey Billington (Chair)

Ksenia Gnevsheva (Secretary)

Shunichi Ishihara (Treasurer)

Yuko Kinoshita

Catherine Travis

Michael Carne (Student Liason)

Elena Sheard (Web Assistant)

Programme Committee

Rosey Billington

Ksenia Gnevsheva

Shunichi Ishihara

Beena Ahmed

Michael Tyler

Conference Assistants

Shubo Li

Marcel Reverter-Rambaldi

Volunteers

Heba Bou Orm

Dimitrije Karadarevic

Emma Keith

Sonia Kulkarni

Keira Mullan

Thomas Powell-Davies

Gan Qiao

Larissa Schwenke

Will Somers

Conference Award Winners

ASSTA New Researcher Award

Angelo Dian

Yanping Li

Hannah White

Review Panel

Mark Antoniou

Brett Baker

Catherine Best

Sasha Calhoun

Lynn Clark

Josh Clothier

Felicity Cox

Karen Croot

Katherine Demuth

Gerry Docherty

Julien Epps

Janet Fletcher

Paul Foulkes

Helen Fraser

Andy Gibson

Simón Gonzalez

David Grayden

Adele Gregory

Bernard Guillemin

Jonathan Harrington

Mark Harvey

Jen Hay

Yusuke Hioka

Colleen Holt

Rebecca Holt

Vincent Hughes

Solène Inceoglu

Kathleen Jepson

Carmen Kung

Debbie Loakes

Olga Maxwell

Kirsty McDougall

Carmel O'Shannessey

Joshua Penney

Michael Proctor

Phil Rose

Mridula Sharma

Mary Stevens

Tünde Szalay

Marija Tabain

Catalina Torres

Kimiko Tsukada

James Walker

Paul Warren

Catherine Watson

Janice Wing Sze Wong

Ivan Yuen

Eighteenth Australasian International Conference on Speech Science and Technology

SST 2022

Table of Contents

Effects of mobile phone transmission on formant measurements: a large-scale examination based on 306 Japanese male speakers

Yuko Kinoshita[1], Takashi Osanai[2]

[1] College of Arts and Social Science/Asia and the Pacific, The Australian National University
[2]National Research Institute of Police Science, Japan

yuko.kinoshita@anu.edu.au; osanai@nrips.go.jp;

Abstract

This study presents a large scale, well-controlled examination of the effects of mobile phone transmission on the first four formants. We used 306 Japanese male speakers recorded simultaneously with a direct microphone and via a mobile phone network. We found that all four formants were significantly impacted by the mobile phone transmission. Further, we found the impact of mobile transmission was largely unpredictable, and the impacts appear to vary speaker to speaker. This may have significant implications in some application areas, such as forensic phonetics, and also for data collection of speech recorded over phones in general.

Index Terms: formants, mobile phone transmission, Japanese vowels

1. Introduction

The impact of phone transmission on acoustic signals has been of interest in the field of forensic voice comparison (FVC) (e.g. for landlines [1-4], for mobile phone [5-7], and the resulting impact on FVC performance [8, 9]).

For phoneticians, formants have been one of the key acoustic features: the 2011 international survey on forensic voice comparison practitioners (across 15 countries, 36 respondents) reported that 97% conducted some form of formant analysis [10]. While the field is increasingly shifting to techniques based on automatic speaker recognition [11-17], the capacity of formants to link acoustic information to vocal tract configurations is still attractive, as it enables incorporation of linguistic knowledge into interpretation and analysis. It is impossible to gain a full understanding of how, and to what extent, various linguistic and non-linguistic factors affect speech acoustics. However, formants can provide a pathway to meaningful interpretations of a subset of acoustic information.

In this light, we conducted a large scale, well controlled study on the impact of mobile phone transmission (the 'mobile phone effect') on formants. Pre-existing studies are based on populations too small for detailed statistical examination. This study aims to fill this gap by comparing formants extracted from the five Japanese vowels uttered by 306 male speakers of Japanese (40-44 utterances for each vowel) in the NRIPS database [18]. They were simultaneously recorded through two different conditions: a direct microphone; and transmission through a Japanese mobile phone network, which makes this database ideal for our purpose. We extracted the first four formants (F1, F2, F3, and F4), using a semi-automatic approach. We developed this approach to handle the scale of this study: it simulates a decision-making process which a human formant measurer would take in formant selection and corrections, as described in the methodology section.

In this study, we have three objectives. The first is to produce a large formant dataset on two recording conditions, as we believe such data are by themselves useful information.

The second is to examine tendencies in the mobile effect on formants. A previous study on English speakers (six male and six female) compared direct microphone recordings and those transmitted via a mobile network. It reports some concerning results: mobile transmission can significantly shift F1 upwards, especially with high vowels. A majority of speakers had a 20-30% rise in F1, and one male and one female had a 50% and a 40% rise [5]. They found some impact on F2 and F3, though weaker than on F1. They also found that the severity of the mobile transmission effect was not consistent across speakers. Another relevant study examined how the Adaptive Multi-Rate (AMR) codecs of mobile phones affected formants, using three female and five male Australian speakers. It found quite large effects [7], which depended on the choice of codecs, and the effect was greater with higher formants, unlike what was reported in [5]. The female speakers appeared particularly susceptible to this effect, especially with F2 and F3 – the distributions of the formants processed with AMR codecs were far removed from those from the microphone recording. In another study, spectrogram observations showed that all codecs introduce an area of missing energy in the low frequency regions [6]. They pointed out that these white island effects would influence FFT and LPC analyses, which is concerning as formant detection is based on LPC analyses. The characteristics of mobile phone transmission effects change dynamically in response to network conditions; AMR dynamically switches among one of eight sub-codecs in response to its assessment of the condition of its transmission channel. In forensic casework, however, these conditions are unknown to analysts. The large dataset for this study will enable us to explore overarching tendencies at the endpoint of mobile transmission more reliably than these previous small-scale studies, leading to better informed interpretation of forensic casework speech data.

These observations lead us to our third objective: examining whether the severity of the mobile phone effects is speaker dependent. Although we tend to classify speakers into male and female binary sex categories, in reality there are speakers whose speech possesses acoustic characteristics atypical of their biological sex. Considering how severely the codec affected formants of female speakers, we can reasonably speculate that codecs can impact some speakers more severely, even among an all-male population: perhaps those with physically smaller vocal apparatus. If there is a large variation between speakers in how mobile phone transmission affects formants, it calls for careful consideration of what we mean by "comparable" in conducting linguistic analyses. Is matching the conditions of the recording channel and sex of the speaker sufficient, or do we need to control for other factors which contribute to the severity of the impact? This question can have serious implications in some contexts, such as FVC.

2. Methodology

2.1. Database, speakers, and speech materials

This study used 306 male native speakers of Japanese from the NRIPS database [18]. At the time of recording, they were aged 18–76 years and lived in Tokyo and its vicinity, but with varying native dialectal backgrounds. We chose this database for its availability. In the 15 years since it was created, the relevant technologies have evolved greatly, so our results are not directly applicable to the most recent mobile recordings. However, it is still useful data for exploration, and it can provide useful information for analysts, in casework where old recordings needed to be revisited.

All speakers in the NRIPS database were recorded in two non-contemporaneous recording sessions, two to three months apart. They performed the same recording tasks twice at each recording session, and the whole process was recorded simultaneously through multiple channel settings. From these, we chose direct microphone recordings (labelled Ch1) and the same utterances recorded at the receiving end of a mobile phone network (labelled Ch3). The mobile phone used for the data collection was DOCOMO FOMA NEC N902i and the transmission system was W-CDMA. The database was recorded at a sampling frequency 44.1kHz originally, but was down sampled to 8kHz, as Ch3 does not contain acoustic information beyond this.

We selected read-out (C)V syllables as the target speech material, prioritising reliable extraction of formants. The selected consonantal environments are: ∅ (no consonant), /k/, /s/, /t/, /h/, /r/, /g/, /z/, /d/, /b/, and /p/. They were followed by one of the five vowel phonemes of Japanese, /a/, /e/, /i/, /o/, and /u/. This resulted in 11 different phonological contexts for /a/, /e/, and /o/ and ten for /i/ and /u/, as the pairs of syllabary, ぢ /di/ – じ /zi/ and づ /du/ – ず /zu/, are phonetically merged to [dʑi] and [dzɯ] respectively. Every speaker had exactly the same syllables to read out, guaranteeing equal phonological conditions across speakers.

2.2. Formant extraction process

Manual formant measurement is time-consuming and prone to measurer-dependent variability [19]. However, the alternative – automatic extraction – is known to miss or misidentify target formants. Some areas of phonetics have embraced automatic formant extraction, and some new approaches to improve this have been proposed (e.g. [20]). However, they mostly focus on F1 and F2, presumably for the interest in linguistic information rather than speaker information. F1 and F2 have relatively strong spectral energy and are easier to detect. F3 and F4, where more speaker information lies [21-28], are harder to extract automatically, as these spectral peaks are less salient. Unreliable formant detection requires human interventions and corrections, which could introduce supervisor-dependent measurement variability [29], as well as being time consuming. To overcome this, we developed a systematic and replicable approach, which simulates the process that human measurers would apply, as described below.

Step 1: Human measurers often try to measure formants by identifying a stable section of formant trajectory as the target section. Thus, we postulated that the formants of monophthongs can be reasonably represented by the most typical values of the multiple measurements sampled across the duration of a given vowel. We set the formant analysis range of Praat at 0-4kHz and sampled poles every 0.005 seconds. Although our targets were F1 to F4, we chose to extract five poles, since removing peaks that are not our interest is much easier than systematically searching for missed peaks. We called these poles 'peak1-5', not formants, as they may not actually be the target formants.

Step 2: Human measurers can easily detect and reject outliers based on the continuity of the trajectory. We simulated this by producing a histogram with 100Hz bins for peak 1 measurements and identifying the most populated bin and its immediately adjacent bins on both sides. The frequency range represented by these selected three bins was identified as a likely F1 range for this token. All peak 1 measurements which fell in this range were labelled as the updated peak 1. We processed peak 2 in almost the same way: we gathered peak 2 measurements that were higher than the F1 range. We produced a histogram from them, and identified the potential F2 range in the same way as we did with the F1 range. All measurements within the potential F2 range were classified as updated peak 2. For the tokens which did not have a peak 1 measurement that fitted in the likely F1 range, we checked if its peak 2 measurement did. If so, it was considered as a misidentified case and added to the updated peak 1 data. We applied the same process to the measurements from peaks 3 to 5, except the potential frequency ranges were set wider to reflect their naturally greater variations. We identified the most populated bin in the histogram, and the two adjacent bins on both sides of it were used as the potential frequency range. This resulted in time series lists of the measurements of the updated peaks 1-5, which are located within the likely formant ranges. There were a few situations where the histogram appeared bimodal. In such cases, we applied the same process to both peaks and recorded both, giving up to six peaks.

Step 3: We fitted a kernel density estimation to the lists of updated peak measurements, using the density function of the statistical package R. The density function of R automatically assigns x-coordinates by dividing the distance between the minimum and the maximum values in equidistance (we used default 1024 coordinates). The coordinates of the maximum point and its immediately adjacent points are fitted to a quadratic function, and the maximum value of this function was recorded as the most representative frequency value of the formant of a given vowel.

Step 4: From these five (or six, where the token had an additional peak) values, we selected four that are most likely to be F1, F2, F3 and F4. Human measurers would use their phonetic knowledge of likely frequency ranges for each vowel and formant. To simulate this, we firstly calculated the overall distributions of the five peaks for each vowel from all tokens from all 306 speakers (13,464 tokens for /a/, /e/, and /o/ and 12,240 tokens for /i/ and /u/) based on Ch1 (microphone) data, as then are likely to be less affected by external factors. Then, the most typical values for the first four peaks were set as the initial values for F1, F2, F3 and F4.

Step 5: To assign five peaks to four formants, we have five possible patterns (see Table 1. Note that we had 15 patterns, where six peaks were detected). For each token, we selected the pattern in which the four peaks showed the least distance to each of the initial formant values. After performing this process to all tokens, the population means were calculated, and the initial values were replaced with these temporary mean values. This peak assignment process was repeated to re-examine which of the five patterns are the closest to the model population formant values. This process was repeated until the mean formant values stabilised.

As well as formants, we sampled F0 using Praat with the analysis range set at 75-350Hz, as we postulated that F0 may be

used as a predictor for the severity of the mobile transmission effect. F0 information was included in the statistical modelling.

Table 1. Possible mapping of peak to formants (for the case of five peaks)

	peak1	peak 2	peak 3	peak 4	peak 5
pattern 1	F1	F2	F3	F4	
pattern 2	F1	F2	F3		F4
pattern 3	F1	F2		F3	F4
pattern 4	F1		F2	F3	F4
pattern 5		F1	F2	F3	F4

2.3. Statistical Analysis

The formant data was first examined with descriptive statistics and visualization, and compared to the results of the previous studies. Then, the mobile phone effect was analysed with linear mixed effects modeling, since our data has a large number of recordings from each speaker, which makes each data point not truly independent [30]. Since the effect of mobile transmission is expected to be non-uniform across the frequency range [6], we built separate models for each formant. We set channel difference, F0, and vowel as the fixed effects, and speakers as a random effect.

3. Results

3.1. Descriptive statistics

3.1.1. Overall effect

Table 2 presents the summary statistics from 306 speakers, as well as proportional between-channel differences (Ch3-Ch1 divided by Ch1) in the column 'Diff%'.

Table 2. Summary descriptive statistics

		Ch1 Mean	SD	Ch3 Mean	SD	Ch3-Ch1 Diff %
a	F1	693	84.4	672	57.8	-3.0%
	F2	1259	92.3	1279	132.9	1.6%
	F3	2632	159.7	2553	131.3	-3.0%
	F4	3516	145.6	3268	178.9	-7.0%
e	F1	494	39.7	486	41.6	-1.7%
	F2	1928	127.1	1914	125.6	-0.7%
	F3	2589	142.2	2519	131.8	-2.7%
	F4	3476	139.1	3211	162.6	-7.6%
i	F1	377	36.5	363	36.5	-3.6%
	F2	2154	159.2	2135	135.9	-0.9%
	F3	2939	139.2	2895	161.4	-1.5%
	F4	3426	111.3	3268	112.5	-4.6%
o	F1	488	48.7	492	44.5	0.8%
	F2	911	98.6	882	70.8	-3.1%
	F3	2689	157.2	2612	172.6	-2.8%
	F4	3357	147.2	3256	102.3	-3.0%
u	F1	394	35.0	398	31.4	0.9%
	F2	1392	150.9	1407	143.2	1.1%
	F3	2380	174.3	2385	128.3	0.2%
	F4	3439	150.6	3287	102.9	-4.4%

In [5], F1 of high vowels in the mobile phone recordings was reported to be 29% higher on average than that of the microphone recordings. Overall, our formant data revealed no such tendency; the mean cross-channel differences for F1 were small, and for /i/ Ch3 were marginally lower. The same study also found that low F2 were lifted and high F2 lowered, but this too was not apparent in our data. For the vowel with the lowest F2, /o/, the mobile phone appears to have had a lowering effect. The vowels with relatively high F2, /i/ and /e/, showed very little difference across the two channels. The only consistent effect of the mobile transmission observed is in F4; the mobile transmission lowers them.

Figure 1 presents the 306 speakers' formants separately for each vowel. Formants extracted from the same utterances but recorded in two different channels (Ch1 and Ch3) were plotted against each other. If the channel difference had no effect on formants, the datapoints should fall very close to the diagonal lines, but that is not the case here. Although the summary statistics did not reveal marked differences between the two channels, the mobile phone transmission appears to have had considerable impact on the formants — and worse still, in mostly unpredictable ways. This suggests difficulties for reliable channel compensation with formants.

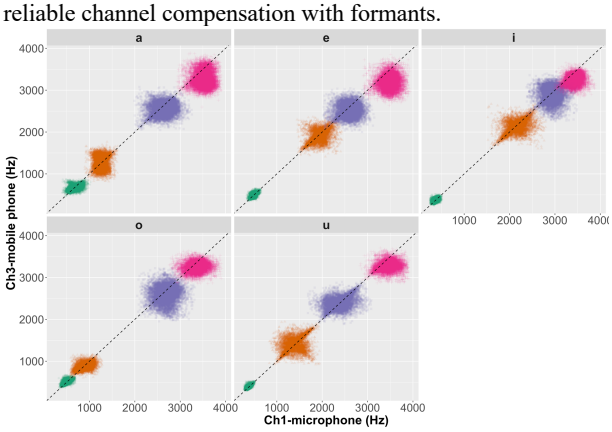


Figure 1. F1- F4 extracted from the same utterances recorded via Ch1 and Ch3 plotted against each other.

3.1.2. Between-speaker variability of the impact

Next, we examined how the mobile transmission effect differed between speakers. Figure 2 presents two example speakers, whose mean channel difference was the smallest and the largest among the 306 speakers. Initial visual inspection seems to support our supposition: that mobile transmission does not impact everyone in the same way or to the same extent.

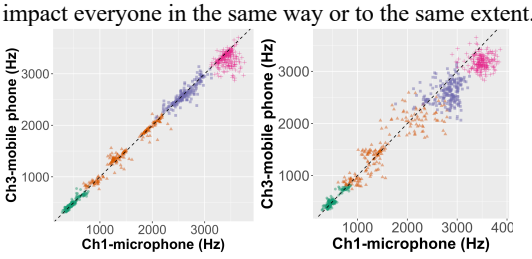


Figure 2. Example of speaker dependency of the mobile effect

3.2. Linear mixed effect model analysis

These initial observations warranted further examination, so we conducted linear mixed effects model analysis. In this section, we examine the between-speaker variability of the mobile transmission effect first, as it determines the better model for overall analysis. We built two models with different slope settings for the random effect (speakers): Model 1 had a fixed slope; Model 2 had a varying slope in relation to the channel

effect. Model 2 reflects our hypothesis that the mobile transmission effects on formats are, at least to some degree, speaker dependent. Using the lmer() function of the R lme4 package, each formant was modelled as below:

1. Hz ~ Channel + F0 + Vowel + (1 | Speaker)
2. Hz ~ Channel + F0 + Vowel + (1 + Channel | Speaker)

We then proceeded to test the fit of these models using the R anova() function. Model 2, which allows the slope of the random effect to vary, clearly better performed with all four formants (Table 3). This indicates that different speakers were impacted differently by the mobile transmission.

Table 3. Comparison of the two models.

		AIC	BIC	Chisq	Pr
F1	Model1	1348391	1348479		
	Model2	1346957	1347064	1438.3	< 2e-16
F2	Model1	1605298	1605386		
	Model2	1604810	1604917	492.36	< 2e-16
F3	Model1	1656963	1657051		
	Model2	1654132	1654240	2835.3	< 2e-16
F4	Model1	1642606	1642694		
	Model2	1635885	1635992	6725.4	< 2e-16

Next, we examined the fixed effects based on Model 2: channel, F0, and vowel. Table 4 presents the summary; the columns 'Est', 'SE', 'df', 't', and Pr(>|t|) denote estimate, standard error, degree of freedom, t-value, and p-value, respectively. Our primary interests here are channel effect ('Ch') and F0. 'Int' denotes intercept. We see that all four formants were affected by the mobile phone effect. F0 appeared to strongly predict F1, and less so F2, and had no discernible relationship with F3 and F4. Formants, especially F1 and F2, are closely linked to articulatory gestures. Each vowel has different gestures resulting in different formant frequencies. Vowels as fixed effects are, thus, not of interest in this analysis and are omitted from Table 4. However, we note that the vowel effects were found to be significant for F3 and F4 as well, which was somewhat surprising, as they are generally considered to reflect more speaker information than linguistic information.

Table 4. Fixed effect results

		Est	SE	df	t	Pr(>|t|)
F1	Int	659.9	1.76	1058	374.77	<2e-16
	Ch	-7.3	0.67	305	-10.84	<2e-16
	F0	0.2	0.01	76400	22.19	<2e-16
F2	Int	1278	4.3	1467	297.16	<2e-16
	Ch	-5.61	1.28	305	-4.38	1.67E-05
	F0	-0.05	0.02	52370	-2.03	0.0424
F3	Int	2619	5.27	1225	497.11	<2e-16
	Ch	-53.92	2.8	305	-19.25	<2e-16
	F0	0	0.03	17550	0.17	0.864
F4	Int	3486	4.84	1359	720.17	<2e-16
	Ch	-185.7	3.84	305	-48.33	<2e-16
	F0	-0.01	0.03	25330	-0.42	0.673

Figure 3 presents how speakers' formant values were shifted by transmission through the mobile network. For better visibility, we plotted a randomly selected group of 55 speakers, not 306. Each line represents a different speaker. It shows considerable speaker variations in both the direction and the extent of the impact, suggesting that the overall tendency across the population would not serve well in predicting the mobile transmission effects on an individual's speech recordings. Two possible reasons can be put forward: speakers' spectral characteristics are shaped by their vocal tract configurations, and the impact of switching codecs. However, we speculate that the latter is less likely, as it is reasonable to assume that each speaker's utterance would have been processed with similarly varying codecs.

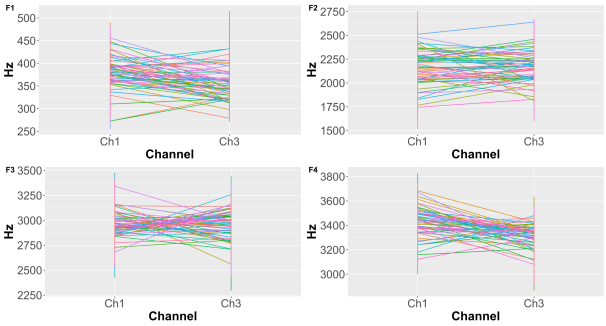


Figure 3. Example of speaker dependency of the mobile phone effect, taken from 55 speakers /i/ vowel.

4. Discussion

These findings point to potentially serious implications. Firstly, use of formants in FVC needs more caution, where it requires comparisons between microphone recordings and recordings made via a mobile phone network. The obtained formants may contain additional and largely unpredictable variability even if the two recordings originated from the same speaker. This increased within-speaker variability makes it less likely for the assessments to result in strong support for a same-speaker hypothesis even when that is factually correct.

Selection of development data for calibration and validation of FVC systems may also require rethinking. A development dataset needs to have 'similar' characteristics to the forensic samples. Gender and channel conditions together with linguistic variety were commonly considered in selecting development data, but our findings suggest this may be inadequate. Further exploration into how we define 'similar' in speech characteristics is in due. Such conditions would also depend on the acoustic features and communication technology in use.

Further, in other types of linguistic research too, phone speech may feature, as it is so ubiquitous. Some caution may be due where this is analysed acoustically.

5. Conclusion

This study performed large scale formant extraction of five vowels from Japanese speech, using an original semi-automatic approach. Through a well-controlled experiment, it revealed that mobile phone transmission affected the first four formants. Closer analysis uncovered a far more complex problem: the mobile phone effect varied across speakers, and its size and direction are difficult to predict. This suggests that the commonly used categories—gender and channel conditions—are inadequate for selecting development data in FVC and further research is needed.

6. Acknowledgement

We thank our anonymous reviewers for their insightful and helpful comments.

7. References

[1] A. Hirson, P. French, and D. Howard, "Speech fundamental frequency over the telephone and face-to-face: some implications for forensic phonetics," in *Studies in General and English Phonetics*, J. W. Lewis Ed. London: Routledge, 1995, pp. 230-240.

[2] H. J. Künzel, "Beware of the 'telephone effect': the influence of telephone transmission on the measurement of formant frequencies," *Forensic Linguistics* vol. 8, no. 1, pp. 80-99, 2001.

[3] S. Lawrence, F. Nolan, and K. McDougall, "Acoustic and perceptual effects of telephone transmission on vowel quality," *International Journal of Speech, Language & the Law,* vol. 15, no. 2, 2008.

[4] P. J. Rose, D. Lucy, and T. Osanai, "Linguistic-acoustic forensic speaker identification with likelihood ratios from a multivariate hierarchical effects model: A "non-idiot's bayes" approach," in *the 10th Australian International Conference on Speech Science & Technology*, Sydney, S. Cassidy, Ed., 8-10/12/2004 2004: Australian Speech Science and Technology Association, pp. 402-407.

[5] C. Byrne and P. Foulkes, "The 'mobile phone effect' on vowel formants," *International Journal of Speech Language and the Law,* vol. 11, no. 1, pp. 83-102, 2004.

[6] B. J. Guillemin and C. Watson, "Impact of the GSM mobile phone network on the speech signal: some preliminary findings," *International Journal of Speech, Language & the Law,* vol. 15, no. 2, 2008.

[7] B. J. Guillemin and C. Watson, "Impact of the GSM AMR speech codec on formant information important to forensic speaker identification," in *Proceedings of the 11th Australian International Conference on Speech Science & Technology*, 2006, pp. 483-488.

[8] A. Alexander, D. Dessimoz, F. Botti, and A. Drygajlo, "Aural and automatic forensic speaker recognition in mismatched conditions," *The International Journal of Speech, Language and the Law,* vol. 12, no. 2, pp. 214-234, 2005.

[9] C. Zhang, G. S. Morrison, E. Enzinger, and F. Ochoa, "Effects of telephone transmission on the performance of formant-trajectory-based forensic voice comparison – Female voices," *Speech Communication,* vol. 55, no. 6, pp. 796-813, 7// 2013, doi: http://dx.doi.org/10.1016/j.specom.2013.01.011.

[10] E. Gold and P. French, "An international investigation of forensic speaker comparison practices," in *Proceedings of the 17th International Congress of Phonetic Sciences, Hong Kong, China*, 2011, pp. 1254-1257.

[11] A. P. Ajit, A. George, and L. Mary, "I-Vectors for Forensic Automatic Speaker Recognition," in *2018 International CET Conference on Control, Communication, and Computing (IC4)*, 2018: IEEE, pp. 284-287.

[12] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust dnn embeddings for speaker recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018: IEEE, pp. 5329-5333.

[13] M. Jessen, J. Bortlík, P. Schwarz, and Y. A. Solewicz, "Evaluation of Phonexia automatic speaker recognition software under conditions reflecting those of a real forensic voice comparison case (forensic_eval_01)," *Speech Communication,* vol. 111, pp. 22-28, 2019/08/01/ 2019, doi: https://doi.org/10.1016/j.specom.2019.05.002.

[14] M. Jessen, G. Meir, and Y. A. Solewicz, "Evaluation of Nuance Forensics 9.2 and 11.1 under conditions reflecting those of a real forensic voice comparison case (forensic_eval_01)," *Speech Communication,* vol. 110, pp. 101-107, 2019/07/01/ 2019, doi: https://doi.org/10.1016/j.specom.2019.04.006.

[15] F. Kelly, A. Fröhlich, V. Dellwo, O. Forth, S. Kent, and A. Alexander, "Evaluation of VOCALISE under conditions reflecting those of a real forensic voice comparison case (forensic_eval_01)," *Speech Communication,* vol. 112, pp. 30-36, 2019/09/01/ 2019, doi: https://doi.org/10.1016/j.specom.2019.06.005.

[16] G. S. Morrison and E. Enzinger, "Multi-laboratory evaluation of forensic voice comparison systems under conditions reflecting those of a real forensic case (forensic_eval_01) – Conclusion," *Speech Communication,* vol. 112, pp. 37-39, 2019/09/01/ 2019, doi: https://doi.org/10.1016/j.specom.2019.06.007.

[17] J. Rohdin *et al.*, "End-to-end DNN based text-independent speaker recognition for long and short utterances," *Computer Speech & Language,* vol. 59, pp. 22-35, 2020/01/01/ 2020, doi: https://doi.org/10.1016/j.csl.2019.06.002.

[18] H. Makinae, T. Osanai, T. Kamada, and M. Tanimoto, "Construction and preliminary analysis of a large-scale bone-conducted speech database," (in Japanese), *IEICE technical report,* vol. Speech 107, no. 165, pp. 97–102, 2007. [Online]. Available: http://ci.nii.ac.jp/naid/40015600747/.

[19] M. Duckworth, K. McDougall, G. de Jong, and L. Shockey, "Improving the consistency of formant measurement," *International Journal of Speech, Language & the Law,* Article vol. 18, no. 1, pp. 35-51, 2011, doi: 10.1558/ijsll.v18i1.35.

[20] K. Evanini, S. Isard, and M. Liberman, "Automatic formant extraction for sociolinguistic analysis of large corpora," in *Tenth Annual Conference of the International Speech Communication Association*, 2009.

[21] F. Clermont and P. Mokhtari, "Acoustic-articulatory evaluation of the upper vowel-formant region and its presumed speaker-specific potency," in *Fifth International Conference on Spoken Language Processing*, 1998.

[22] S. Furui and M. Akagi, "Perception of voice individuality and physical correlates," *音響学会聴覚研資,* pp. H 85-18, 1985.

[23] U. G. Goldstein, "Speaker‐identifying features based on formant tracks," *The Journal of the Acoustical Society of America,* vol. 59, no. 1, pp. 176-182, 1976.

[24] X. Lu and J. Dang, "An investigation of dependencies between frequency components and speaker characteristics for text-independent speaker identification," *Speech communication,* vol. 50, no. 4, pp. 312-322, 2008.

[25] P. Mokhtari and F. Clermont, "Contributions of selected spectral regions to vowel classification accuracy," in *Third International Conference on Spoken Language Processing*, 1994.

[26] L. C. Pols, H. R. Tromp, and R. Plomp, "Frequency analysis of Dutch vowels from 50 male speakers," *The journal of the Acoustical Society of America,* vol. 53, no. 4, pp. 1093-1101, 1973.

[27] S. Saito and F. Itakura, "Personal characteristics of the frequency spectrum for vowels," *Annual Bulletin of the Research Institute of Logopedics and Phoniatrics,* vol. 16, pp. 73-79, 1982.

[28] K. N. Stevens, "Sources of inter- and intra-speaker variability in the acoustic properties of speech sounds," *Proceedings of the Seventh International Cons. Phonetic Sciences,* pp. 206-232, 1971.

[29] C. Zhang, G. S. Morrison, E. Enzinger, and F. Ochoa, "Laboratory Report: Human-supervised and fully-automatic formant-trajectory measurement for forensic voice comparison–Female voices," *FVC, EE&T, UNSW Laboratory Report,* 2012.

[30] R. Baayen, "Analyzing Linguistic Data: A Practical Introduction to Statistics Using R. Cambridge University Press," 2008.

Likelihood Ratio-based Forensic Semi-automatic Speaker Identification with Alveolar Fricative Spectra in a Real-world Case.

Phil Rose

Independent Researcher, Australian National University Emeritus Faculty

https://philjohnrose.net

Abstract

A real-world forensic voice identification is described in a case involving the blowing-up of a car, three suspects, and a miniscule amount of speech evidence. Necessary stages in the estimation of a likelihood ratio are described, based on the bandlimited cepstral spectral acoustics of two alveolar fricatives /s/ and /z/ in the questioned utterance. The speech evidence is shown to be very much more likely assuming one of the suspects said the utterance.

Index Terms: Forensic voice comparison, alveolar fricative, Bayes' theorem, bandlimited segmental cepstrum, validation.

1. Introduction

One evening in 2017 three young men, K M and C, drove to the Australian town of Wagga Wagga where, with apparently no more sinister motive than amusement, they blew up a stationary unoccupied car belonging to an acquaintance. With a view to posting on social media (!), the incident was videoed from within their car, with verbal commentary, on a mobile phone. At the time the recording starts, it was agreed that two of the three were inside their stationary car looking on, and the offender was outside laying the timed explosive device. The offender then rejoined the other two and the video captures the explosion through the rear window as they drive away. The audio contains recordings of several short utterances, both before and after the time the offender returned to the car.

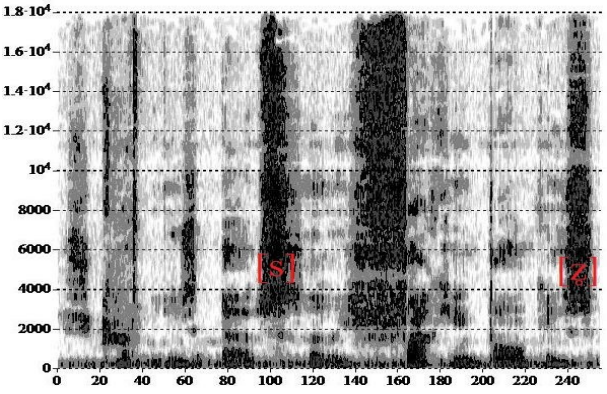


Figure 1: *Spectrogram to 18k of questioned utterance* this cunt is absolutely fucking bonkers *showing durational and spectral extent of noise from alveolar fricatives* [s] and [z̪]. *X-axis = duration (csec.), y-axis = frequency (Hz).*

When questioned by police, each participant denied laying the explosives. C maintained he had sat in the rear seat of the car making and commenting on the video. M said he had sat silent in the front while K made the video. K said he sat in the front seat, watching C make the video. The Wagga police specifically wanted to know, of course, whose was the voice on the video before the offender returned the car.

In addition to the fact that the aim is to test exoneration rather than inculpation, this case is unusual and worth documenting for several other reasons. Because the questioned speaker was one of three, an initial estimate of the prior probabilities is easy and the mandating authority can be instructed as to how Bayes' theorem can be used to estimate a posterior probability. The case also shows how – sometimes! – forensic voice comparison is possible with a miniscule amount of questioned data backed up by phonetic knowledge. The parametrisation, with a bandlimited segmental cepstrum, involved a nice demonstration of the power of combining acoustic phonetics and signal processing. Finally, it shows the crucial element of validation – demonstrating that your system actually does what you claim it to do. All the case details may be found in the anonymised full report at [1].

2. Questioned Data

Logically, the only material relevant to the voice comparison were two short utterances recorded in the absence of the offender, before he returned to the car (there were no utterances after the offender returned to the car which were incriminating by content). The second utterance was said *sotto voce* and was of no use. The first utterance – *This cunt is absolutely fucking bonkers* – contained no usable vocalic material. Likelihood ratio-based testing in [2] showed all its stressed vowels – /ae/ and / ʉː/ in *absolutely*, /a/ in *cunt* and *fucking*; and /o/ in *bonkers* – to have rather poor expected evidential strength, with equal error rates ranging from 25% (/æ/) to 36.6% (/o/). However, previous research into the speaker-identification potential of fricatives with various parametrisations [3-8] suggested a forensic voice comparison might be feasible using the spectra of /s/ and /z/ in *absolutely* and *bonkers*. (Subsequent work [9-12] has confirmed this).

Figure 1 shows a 60dB dynamic range wide-band spectrogram to 18.5 kHz of *This cunt is absolutely fucking bonkers*, which can be seen to last for about 2.5 seconds. The /s/ is at ca. csec. 100; the /z/, which shows typical word-final devoicing, at ca. csec. 240. Both last for about 10 centiseconds. The high intensity noise centered at csec. 150 is from an emphatically produced /f/ in *fucking*. The fricative energy extends to about 18 kHz, and well-defined vocalic formant structure can be seen below 4 kHz.

Figure 2 shows the questioned alveolar spectral acoustics (FFT and 14th order LPC) to 8 kHz. The tokens are unremarkable, showing spectral properties predicted from the acoustic theory of speech production [13 pp.379-389, 398-403]. The peak between 3 and 4 kHz is the quarter-wavelength resonance of the front cavity, the peak around 6 kHz is the half-wavelength resonance of the constriction. Even the back cavity resonance is visible just below 2 kHz, attesting to the quality of the recording.

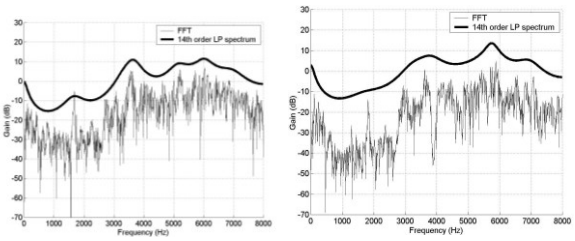


Figure 2: *Spectral acoustics of questioned /s/ (left), and /z/. X-axis = frequency (Hz.), y-axis = gain (dB).*

3. Suspect Data

Recordings were available from C K and M during their police interviews. M's recording was audibly slightly echoic in places, otherwise, for once!, the recordings were of rather good quality. Tokens of /s/ and /z/ in the suspects' speech during their police interviews were identified. To preserve comparability with the questioned tokens, suspects' tokens in the vicinity of rounded vowels were omitted to avoid coarticulation effects from lip-rounding, which will in particular lower the resonance associated with the front cavity. Because of the well-known utterance-final conditioning of duration and voicing of voiced fricatives in English, only utterance-final tokens of /z/ were selected. One of the forensically useful aspects of /s/ and /z/ is that they occur frequently in speech: the recordings of C K and M contained 48, 43 and 65 useable tokens of /s/, and 21, 34 and 16 useable tokens of /z/ respectively.

4. Processing

Suspects' /s/ and /z/ tokens were extracted and saved with *Praat*. Figure 3 is an example of one of C's prepausal /s/ tokens showing the high frequency noise portion saved. As can be seen, the broadband energy between about 3 and 7 kHz lasts for about 20 centiseconds.

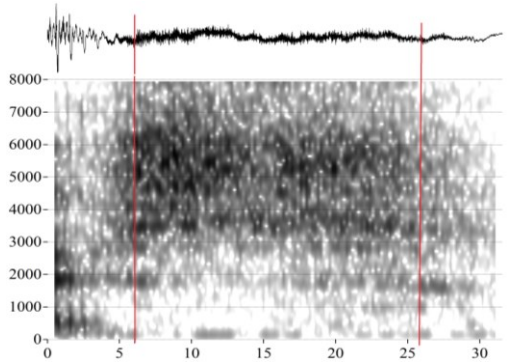


Figure 3: *Spectrogram and waveform of a prepausal /s/ token showing portion of high frequency energy associated with the fricative noise. X-axis = duration (csec.), y-axis = frequency (Hz)*

The tokens were then further processed in Matlab, where a set of eight linear prediction cepstral coefficients was extracted. A hamming widow was applied over the whole of the fricative: in this way one obtains a good estimate for the central portion of the noise. This so-called *segmental cepstrum* procedure smooths the spectrum to an extent which facilitates forensic voice comparison [6 8 10 11]. Figure 4 shows the resulting 8th order cepstral spectra of the individual /s/ and /z/ tokens of each of the three speakers (in blue), and their mean cepstral spectrum (in red). The cepstral spectra of the questioned /s/ and /z/ are plotted in black. Figure 4 shows some clear spectral differences between the speakers. With a prominent peak at about 4 kHz, K's spectra are the most different. This may be related to the fact that he sometimes whistled his /s/ and /z/. This peak will mean that K will have a relatively large second cepstral coefficient. C and M's spectra are more similar, but still differ in amplitude range, with C having a greater range than K. This will be reflected in the amplitude of the first cepstral coefficient: C's will be bigger. The questioned spectra are visually closer to C than to K and M.

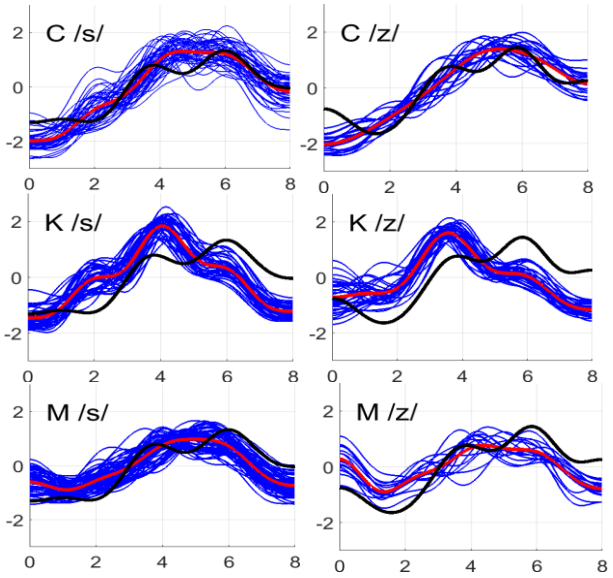


Figure 4: *Cepstral spectra for known and questioned tokens. Blue = individual known tokens, red = known mean, black = questioned. X-axis = frequency (kHz.), y-axis = gain (dB).*

5. Evaluation of Evidence

We know the questioned utterance (U*q*) comes from one of three speakers, thus three separate hypotheses – H_C H_K H_M – must be evaluated in this case: that U*q* was said by suspect C, K and M respectively. This is, of course, a classic identification task, but from a forensic point of view its solution is not simply a question of finding the speaker who has the closest acoustics to U*q*. As is by now well-known, the legally and logically correct approach must be to estimate the strength of evidence, *aka* likelihood ratio, for the three hypotheses [14 15]. The likelihood ratio (LR) is the ratio of the probabilities of the evidence under competing hypotheses: it quantifies how much more likely you are to get the evidence under one hypothesis than under an alternative. If you have the strength of evidence and know the probability of the hypothesis before that evidence is adduced, it is possible to estimate, with Bayes' theorem, what everyone wants to know: the probability of the hypothesis, given the evidence.

With all eight cepstral coefficients the results were pretty unequivocal. The LR for /s/ assuming the unknown speaker was C was 29.6; for K and M it was 6e-09 and 6e-04 respectively, indicating that the questioned /s/ spectrum is far more likely assuming it had come from C than the other two. However, for several reasons these results probably overestimate the strength of evidence that can reasonably inferred. These reasons are addressed below.

6. Validation

The first step in estimating the strength of evidence is validation: determining how reliable one's system is in delivering an accurate LR. The question is especially obvious in this case. After all, just how reliable can a system actually be that uses comparisons involving *only a single questioned [s] and [z] token*?! Validation is, or should be, an essential part of any forensic case work [16 17].

Validation is done by the simple, time-honoured method of seeing what happens with *known* data in circumstances as close as possible to those of the actual case. The car recordings could not be used as there was of course no indication of who said what. Therefore, the suspects' police interviews were used and for each of the three speakers, one of their [s] tokens was extracted in turn to serve as the single questioned datum, and its LR estimated as the ratio of two values: (1) the probability of getting the token's cepstral spectrum assuming it had come from the given speaker (i.e. a known same-speaker comparison); and (2) the probability of getting the token's spectrum assuming it had come from either of the other two speakers (i.e. a known different-speaker comparison). Some additional preprocessing was required to improve the accuracy of the strength of evidence estimate. This is now described.

6.1. Bandlimiting

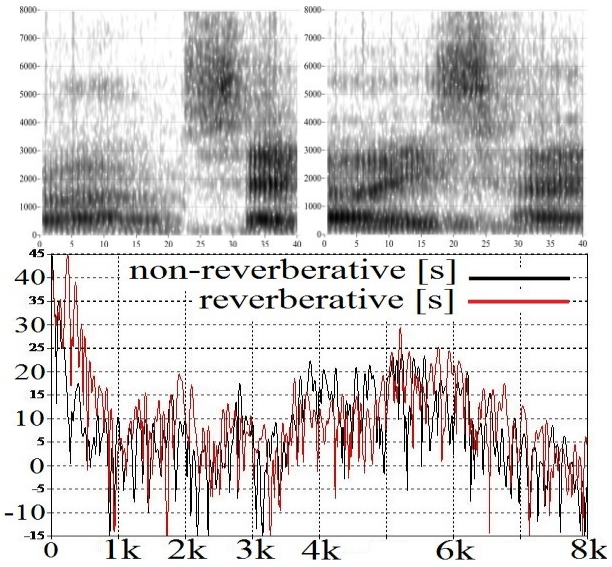


Figure 5: Top = *Spectrograms to 8 kHz of two of M's [s] tokens illustrating possible reverberation in right panel. Bottom = FFT of reverberative and non-reverberative [s] tokens. X-axis = duration (csec.)*

As already mentioned, although the police recording of C and K was of good quality, M's police recording was audibly more echoic. This has a potentially deleterious effect on the spectrum which should be controlled for. Figure 5 illustrates this. Its top left panel shows a spectrogram of one of M's [s] tokens (in *I said*) that had no clear reverb. The top right panel shows a spectrogram of one of his [s] tokens (in *me saying*), showing reverberation from at least the F1 of the preceding [i] in *me* continuing into the [s]. The bottom panel compares the FFT spectra of the two [s] tokens, where it can be seen that the token with putative [i] F1 reverberation appears to have higher energy over about the first kilohertz. This needs to be controlled for, otherwise it might make M's /s/ tokens appear more different from the questioned recording than he actually is, thus favouring identification with C or K. Therefore, firstly, /s/ tokens from M were chosen that sounded minimally reverberative. Secondly, comparisons were done using cepstral coefficients extracted over the range between 1 and 8 kHz, thus removing the spectral portion below 1 kHz where reverberation might have the greatest effect. These so-called *bandlimited cepstral coefficients* (blccs) allow for parametric specification of any cepstral sub-band within the Nyquist interval [18 - 21]. They have great potential in forensic voice comparison because they allow one to focus on the frequency ranges suspected to contain the most speaker-specific information.

6.2. Dimensionality

Estimating the likelihood ratio with a segmental cepstrum involves estimating the probability density of multivariate data. The accurate estimation of a probability density depends crucially on the number of observations, and the more dimensions, the more observations are needed to be sure that all parts of the high dimensional distribution are adequately sampled. In order to adequately sample a 7-dimensional object one would require a hair-raisingly high number of about 43,700 /s/ tokens [22 p.94]. The highest dimensionality supported with the available sample size – between about 40 and 65 /s/ tokens – is three [22 p.94]. This means that the safest choice is one based on comparison with three blccs, and only the first three were used.

6.3. Calculation and calibration of LR-scores

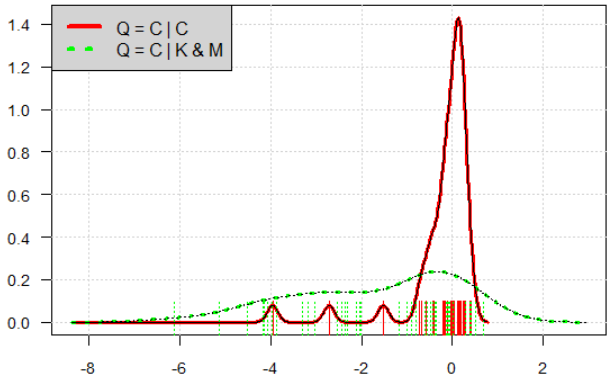


Figure 6: *Kernel density probability density distributions and rugplots for known same-speaker (red) and different-speaker (green) $\log_{10}$ LR-scores for /s/ spectrum validation assuming C is suspect. X-axis = $\log_{10}$LR, y-axis = probability density.*

As cepstral coefficients are distributed normally and the data is multivariate, LR-scores were estimated using *R*'s *mvdnorm*. (*LR-score* is a useful term for an uncalibrated measure of distance that takes both similarity and typicality into account). Thus for each /s/ token of a given speaker, the multivariate probability of its three blccs was estimated assuming it had come from that speaker, and assuming it had come from either of the other two. The LR-score was the $\log_{10}$ ratio of these multivariate conditional probabilities. Given the number of known tokens, this involved upwards of 40 same-speaker and 1560 different-speaker comparisons for /s/, and upwards of 16 same-speaker and 120 different-speaker comparisons for /z/. Figure 6 shows the results of this process treating C as the suspect. The red line (its peak centered

around 0) indicates the probability density distribution for same-speaker LR-scores, i.e. when single questioned /s/ tokens from C were compared with C's distribution. The green line indicates the probability density distribution for the different-speaker LR-scores, i.e. when single questioned /s/ tokens from C were compared with the distributions from K & M. It can be seen that the same-speaker LR-scores are generally bigger than the different-speaker LR-scores. In particular, the probability of getting a $\log_{10}$LR-score between about +/- 0.5 is very much greater if the token had come from C than from the other two. It is also obvious, however, that there is considerable overlap, with counterfactual LRs. There are even two different-speaker comparisons which have LR-scores actually higher than any of the same-speaker comparisons! This variability, perhaps, is to be expected when you are only looking at a single token, rather than the mean of several, but the reality of the case is such that only a single token is available, and so the validation has to reflect that.

The same-speaker and different-speaker LR-scores were then calibrated to convert them to likelihood ratios [23 24]. Two different calibrations were applied. One used the PAV algorithm in the *Focal* tool-kit [25]. In the other, the log-odds of the same-speaker and different-speaker scores were estimated over the range of LR scores. LRs could then be estimated from the kernel density of the log odds.

Figure 7 shows the resulting Tippett plot for the two types of calibration in comparisons where C is the suspect. The plot for the uncalibrated data is also shown. The likelihood-ratio cost metric Cllr for quantifying the validity of a LR-based detection system [26] reflects how much the system is capable of changing the user's prior belief at the most advantageous prior of 1:1. The Cllr for the uncalibrated system (thin brown line) is greater than unity, indicating that it does not reduce the user's uncertainty. The Cllrs for the calibrated data, on the other hand, are 0.77 (PAV) and 0.58 (log-odds kernel density). This indicates that, with comparisons involving C as the suspect, a calibrated system with just a single questioned /s/ token is clearly capable of giving useful information. It can be also seen that the system has about a 20% equal error rate, comprising an error rate of about 5% for same-speaker comparisons and 35% for different-speaker comparisons.

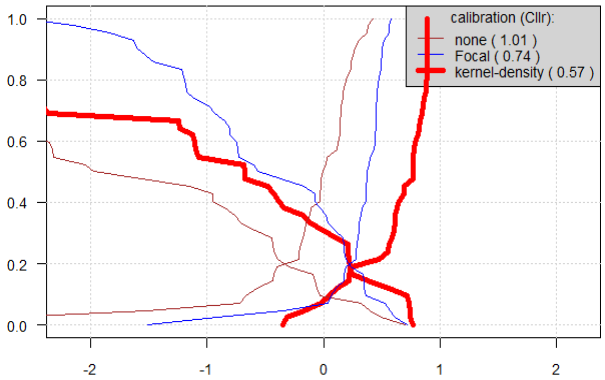


Figure 7: *Tippett plots and associated Cllrs for uncalibrated and calibrated systems assuming C is suspect. X-axis = Log_{10}LR greater than … (different-speaker comparisons) ~ $\log_{10}$LR smaller than … (same-speaker comparisons). Y-axis = cumulative proportion of comparisons.*

Given the validation results, it is then possible to estimate the LR for the actual questioned data against the calibrated distributions. Table 1 shows the final calibrated LR values for both questioned /s/ and /z/. It can be seen that the questioned /s/ and /z/ spectra are far more likely assuming the questioned utterance had come from C than the other two. The enormous value for the /z/ LR under H_C is worth noting. It is because there was rather good separation between his same-speaker and different-speaker validation LR-scores which resulted in Cllrs of 0.34 and 0.32 for both calibration methods. Evidently there was more speaker-specificity hidden in the three blccs of the [z̪] spectrum than the [s].

Table 1. Final calibrated *LR estimates for Questioned /s/ and /z/. H = hypothesis*

H	**/s/**	/z/
H_C	14	332
H_M	6e-05	0.03
H_K	-inf	5e-11

We are left with the problem of how to combine the very different strengths of evidence from /s/ and /z/. Their LRs cannot be combined à la naïve Bayes, as they are sure to be highly correlated. Neither can they be fused as they do not constitute multivariate data. The solution – a crude one – was to re-validate and recalculate with pooled /s/ and /z/ data. This gave a calibrated LR value of ca. 25 for H_C.

7. Posterior Probabilities and Outcome

Normally, the expert is not privy to the prior probability and so cannot logically combine it with their LR to estimate the posterior. They can, however, show the mandating authority how it is done: an example from a Spanish forensic voice comparison case is at [27]. This case is no exception. Even though there are only three individuals who could have said the questioned utterance, the prosecution may have grounds for assuming other than equal priors of 33.3% each. If, for example, they considered K highly unlikely to have said it, their priors might have been more like 45% each for C and M, and 10% for K. With equal priors, and a LR of 25, however, the police could be shown that the posterior odds would be (25/1 * 1/2 =) 12.5 / 1, giving a probability of {12.5 / (12.5 + 1) = } ca. 93% that C was the questioned speaker, thus narrowing it down to either K or M as the perpetrator. For reasons known only to them, the police charged M, not K. He waited until just before start of trial to confess that he had indeed set the bomb.

8. Summary

This paper has described the processing required in applying the LR framework to a real case in an example of what is now called *forensic semi-automatic speaker recognition* [28]. There was exceedingly little evidence: about 20 centiseconds. But otherwise atypically favorable conditions conspired to allow an outcome of use to the mandating authority. Three things are worth emphasising: the forensic potential of voiceless fricative spectra, especially when quantified cepstrally; the complex processing required to achieve a demonstrably valid LR estimate; and the sheer luck of having suspects who differed sufficiently in their speech acoustics for the same sound. It is a pity that not all cases are like this, but that does not excuse us from using the likelihood ratio framework, which, by dint of its explicitness, may very well actually make it possible to do such case-work in the first place.

9. Acknowledgements

Many thanks to my three anonymous reviewers for taking their time to help improve the clarity of this paper. They made some extremely useful suggestions, nearly all of which I was able to incorporate. I also owe thanks to Frantz Clermont both for providing the Matlab script for the bandlimited cepstrum, and for coming up with the concept in the first place.

10. References

[1] Rose, P., Anonymised forensic report, http://philjohnrose.net/pubs/FVC_pubs/index.html, 2017.

[2] Rose, P., "Forensic Speaker Discrimination with Australian English Vowel Acoustics", *Proc. Intl. Congr. of Phonetic Sciences,* Saarbruecken, 2011.

[3] Wolf, J.J., "Efficient Acoustic Parameters for Speaker Recognition", JASA 51: 2044–2056, 1972.

[4] Nolan, F., *Problems and Methods of Speaker Identification*. Unpublished Dip. Linguistics Dissertation, Cambridge University, 1975.

[5] Hillcoat, T.O. *An Evaluation of Selected Sibilant and Nasal Parameters for use in Forensic Speaker Identification*. Unpublished Masters of Letters Dissertation, University of New England, 1994.

[6] Rose, P., "Forensic Voice Comparison with Secular Shibboleths – a hybrid fused GMM-Multivariate likelihood-ratio-based approach using alveolo-palatal fricative cepstral spectra", *Proc. Int'l Conference on Acoustics Speech & Signal Processing* Prague, 5900–5903, 2011.

[7] Kavanagh, C.M., *New Consonantal Acoustic Parameters for Forensic Speaker Comparison*, Ph.D. thesis, University of York, 2012.

[8] Rose, P., "More is better: Likelihood ratio-based forensic voice comparison with vocalic segmental cepstra frontends**",** Int'l Journal of Speech Language and the Law 20(1): 77-116, 2013.

[9] Lo, J. J. H., "One population, two languages: How does language choice affect /s/peaker di/s/crimination?" IAFPA conference poster, 2018.

[10] Zhang C. 张翠玲 and Ding P.丁盼, 擦音LPC倒谱特征在法庭说话人识别中的应用研究 [A study on the application of fricative cepstral features in forensic speaker recognition.] 中国刑警学院学报 5, 117–121, 2019.

[11] Zhang C. 张翠玲 and Ding, P. 丁盼基于LPC倒谱特征融合的法庭说话人识别方法[Forensic speaker recognition based on fused LPC cepstral coefficients],中国刑警学院 2(5), 117–121, 2020.

[12] Smorenburg, L., and Heeren, W., "The distribution of speaker information in Dutch fricatives /s/ and /x/ from telephone dialogues", JASA , 147: 949–960, 2020.

[13] Stevens, K., Acoustic Phonetics, MIT Press, 1998.

[14] Morrison, G.S., Enzinger, E. and Zhang, C. "Forensic Speech Science", in I. Freckelton and H. Selby [Eds] Expert Evidence 99, Thomson Reuters, 1051-6102, 2018.

[15] Rose, P., **"**Likelihood ratio-based forensic voice comparison with higher level features: research and reality**",** in E. Lleida & L. J. Rodriguez-Fuentes [Eds], Recent Advances in Speaker and Language Recognition and Characterisation, Computer Speech and Language Special Issue, 476-502, 2017.

[16] Holdren, J.P., Lander, E.S. et.al. "Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature-Comparison Methods," Science and technology advisory body to the President of the United States, https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/PCAST/pcast_forensic_science_report_final.pdf, 2007.

[17] Lander, E.S., "Response to the ANZFSS council statement on the President's Council of Advisors on Science and Technology Report", Australian J. Forensic Science, 2017.

[18] Clermont, F. and Mokhtari, P., "Frequency-Band Specification in Cepstral Distance Computation", Proc. Australian Int'l Conf. on Speech Science and Technology, 354-359, 1994.

[19] Khodai-Joopari, M., Clermont, F. and Barlow, M., "Speaker variability on a continuum of spectral sub-bands from 297-speakers' non-contemporaneous cepstra of Japanese vowels", *Proc. Australian International Conf. on Speech Science and Technology*, 505-509, 2004.

[20] Clermont, F., Kinoshita, Y. and Osanai, T., "Sub-band cepstral variability within and between speakers under microphone and mobile conditions: A preliminary investigation", *Proc. Australasian Int'l Conf. on Speech Science and Technology*, 317-320, 2016.

[21] Clermont, F. "Linear transformation from full-band to sub-band cepstrum", *Proc. 18th Australasian International Conf. on Speech Science & Technology*, Canberra, 2022.

[22] Silverman, B.W., Density Estimation for Statistic and Data Analysis, Chapman and Hall, 1986.

[23] Morrison, G.S., "Tutorial on logistic regression calibration and fusion: converting a score to a likelihood ratio", Australian J. Forensic Sci., 25(2), 173-197, 2013.

[24] Ramos, D., "Reliable Support: Measuring Calibration of LRs", Keynote at EAFS, The Hague, http://arantxa.ii.uam.es/~dramos/files/2012_08_22_EAFS_Ramos_keynoteReliableSupport_v4.pptx.pdf , 2012.

[25] Brümmer, N., "Focal Toolkit", http://www.dsp.sun.ac.za/nbrummer/focal

[26] Brümmer, N. and du Preez, J., "Application independent evaluation of speaker detection", Computer Speech and Language IEEE Odyssey 2004 Issue 20(2-3), 230-275, 2006.

[27] Lucena-Molina, J., Gascon-Abellan, M. and Pardo-Iranzo, V., "Technical support for a judge when assessing a priori odds", Law, Probability, Risk 14(2), 147-168, 2015.

[28] 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 fuer Polizeiwissenschaft, 2015.

Addressing Sampling-Frequency Mismatch between Speech Data Sets in a Forensic Voice Comparison

Hanie Mehdinezhad[1], Bernard J. Guillemin[1], Balamurali B T[2]

[1]The University of Auckland, New Zealand
[2]Singapore University of Technology & Design, Singapore
E-mail: *h.mehdinezhad@auckland.ac.nz*
E-mail: *b.guillemin@auckland.ac.nz*
E-mail: *balamurali_bt@sutd.edu.sg*

Abstract

Sampling-frequency (f_s) mismatch between Suspect, Offender, and Background speech data sets in a Forensic Voice Comparison (FVC) are discussed and approaches to correct for this are presented. The Bayesian Likelihood-Ratio (LR) framework is used to express the results of a FVC and Gaussian Mixture Model-Universal Background Model (GMM-UBM) is used to calculate LR values. As appropriate, experiments have been conducted on both tokenized and stream data using Mel-Frequency Cepstral Coefficients (MFCCs) as the speech features. The results show that the best approach to correct for f_s-mismatch between speech data sets is down-sampling of the speech data set/sets at higher f_s to match the speech data set/sets at lower f_s.

Index Terms: Sampling frequency mismatch, Forensic Voice Comparison, Bayesian Likelihood Ratio, Gaussian Mixture Model-Universal Background Model.

1. Introduction

In a typical Forensic Voice Comparison (FVC) scenario, the similarity of an Offender voice sample to a Suspect voice sample and the typicality of an Offender voice sample in reference to a Background population, which is case-specific, are assessed to provide the strength of speech evidence. The results of a FVC are then used in order to assist a court of law to make their final decision. In a real forensic scenario, the Offender speech data could be obtained in a number of ways where there is often little/no control over it. For instance, Offender speech data could be obtained from the recordings of a mobile phone or landline phone conversation. On the other hand, there is normally more control over the speech for the Suspect (often recorded during a police interview) and the Background population (usually recorded under highly controlled recording conditions).

In a real forensic investigation, it could be possible that the speech samples associated with Suspect, Offender, and Background data are not at the same sampling-frequency (f_s). For example, the Offender data may be at a lower f_s compared to that of the Suspect and Background data. But when undertaking automatic-based FVC procedures using, for example, MFCCs, all speech data sets must be at the same f_s. Therefore, it is necessary to determine what strategy should be employed in cases where, for example, the Offender speech data is sampled at a different f_s compared to Suspect and Background speech data. Hence, the focus of this paper is on answering the following research question: If the voice samples associated with the Suspect, Offender, and Background data sets are not at the same f_s (this is a form of mismatch that can be called f_s-mismatch between speech data sets), what strategy should be employed to make them the same? We consider the specific case of Offender data being at a lower f_s compared to that of the Suspect and Background data. Should it be up-sampled to match them, or conversely should the Suspect and Background data be down-sampled?

The Bayesian Likelihood Ratio (LR) framework is used to express the outcome of a FVC and a Gaussian Mixture Model-Universal Background Model (GMM-UBM) is used as the statistical procedure to calculate LRs. The speech database used in this study is the XM2VTS (Extended Multi Modal Verification for Teleservices and Security) that is a multi-modal database containing speech recordings sampled at 32 kHz. Thus, 32 kHz is the highest f_s that could be investigated in this research. Mel-Frequency Cepstral Coefficients (MFCCs) are the speech features used as they are commonly used in the FVC arena.

The remainder of this paper is organized as follows. Section 2 provides an overview of the LR framework, followed by a brief discussion on GMM-UBM. Section 3 describes our experimental procedures for addressing f_s-mismatch between speech data sets. The results of these experiments are presented in Section 4, followed by our conclusions in Section 5.

2. Background Information

2.1. *Likelihood Ratio Framework*

The LR, as the name indicates, is the ratio of two likelihoods that mathematically is calculated as:

$$LR = \frac{P(E|H_p)}{P(E|H_d)} \qquad (1),$$

where $P(E|H_p)$ is the conditional probability of E (the evidence) given H_p (the prosecution hypothesis) and this assesses the similarity between the Suspect and Offender speech samples. $P(E|H_d)$ is the conditional probability of E given H_d (the defense hypothesis) and measures the typicality of the Offender speech samples to a relevant Background population. LR values significantly greater than one support the prosecution hypothesis, LR values significantly less than one support the defense hypothesis, and LR values close to one provide little support either way. The Log-Likelihood-Ratio (LLR) is often computed from the LR, where $LLR = log_{10}(LR)$. The sign of the LLR indicates whether it supports the prosecution (positive) or defense (negative) hypothesis and its magnitude indicates the strength of that support.

2.2. *Overview of GMM-UBM*

The GMM-UBM approach [1] is a common technique used in both Automatic Speech Recognition (ASR) and FVC [2-4]. Normally it requires a large amount of data to build a single Background model, namely a Universal Background Model (UBM) [2]. GMM-UBM was originally designed for data-stream-based analysis, but it has also been used by some researchers in token-based analysis [4]. In order to achieve good FVC performance, the UBM is trained on all Background data pooled across speakers. The probability density function of the UBM is estimated using Gaussian Mixture Models (GMMs), with the Expectation Maximization (EM) algorithm [5, 6] being used to train it. The Suspect model is then built by adapting the UBM towards a better fit of the Suspect speech data using the Maximum a Posteriori (MAP) procedure [2]. A score is then calculated as the ratio of the Suspect and Background probability density function values determined at the Offender data points. The scores are then calibrated and fused to get the LR.

2.3. *Measuring FVC performance / Presenting results*

The performance of a FVC is measured by evaluating its accuracy (i.e., validity) and reliability (i.e., precision) [7, 8]. Accuracy indicates the closeness of the obtained result to the true value of the output. The Log-Likelihood Ratio Cost (C_{llr}) [8, 9] is one of the recommended metrics for assessing this, the lower its value, the better the accuracy. Reliability measures the amount of variation that could be expected in LR values, this arising, for example, from improper modelling of the background statistics due to limited data being available of the specified Background population. The Credible Interval (CI) [8, 9] is a popular metric for evaluating this, and again, the lower its value, the better the reliability.

The results of a FVC experiment are often presented using Tippett plots [10] which represent the cumulative proportion of LLR values for both same-speaker and different-speaker comparisons. In these plots (see Figs. 2 & 3) the solid blue and solid red curves are the same-speaker and different-speaker comparison results, respectively. Since positive LLR values support the prosecution hypothesis and negative values support the defense hypothesis, the further apart the curves (i.e., the blue curve towards the right and the red curve to the left), the better would be the performance of the FVC system and therefore generally the lower the C_{llr}. The dashed lines on either side of these solid curves represent the variation in a particular LLR comparison result (i.e., $LLR \pm CI$). The lower the CI value, the higher the reliability of the FVC system.

3. Experimental Procedure

3.1. *Speech Data Set*

The XM2VTS database includes video sequences, face images, and speech recordings [11]. It contains read speech digitized at 16 bits, sampled at 32 kHz, with low Background Noise (BN) level. The language in the database is English with predominantly a Southern British accent. It contains four recording sessions of 295 subjects (156 males, 139 females) collected over a period of 4 months. Sessions were recorded at one-month intervals and during each session each speaker repeated three sequences of words twice. The first two were random sequences of digits from zero to nine: "zero one two three four five six seven eight nine" and "five zero six nine two eight one three seven four". The last sequence was a sentence: "Joe took father's green shoe bench out". In this research only the recordings of the first two sentences (i.e., random sequences of digits) have been used.

Given that the XM2VTS database contains recordings of read speech and the BN level is low, it is acknowledged that it is not very forensically realistic. However, in support of its use in this study, it does include a large number of speakers with similar accents as well as multiple non-contemporaneous recordings, both aspects being highly important in the FVC arena. Only male speaker recordings have been used in this study. Of the 156 male speakers, only 130 were used. The other 26 speakers were discarded because their recordings were either less audible, or they were judged to have different accents to the rest of the speakers (see [12] for a rationale behind discarding recordings on the basis of dissimilar accent).

For our experiments we then down-sampled the recordings to 8 kHz and then up-sampled again depending upon the scenario we wished to investigate. For data-stream-based experiments, the whole utterance (i.e., the two random sequences of digits) after removing all silence segments was used. For the token-based experiments, two diphthongs /aɪ/ and /eɪ/ and one monophthong /i:/ extracted from the words "nine", "eight" and "three", respectively, were used. Audio editing programs Wavesurfer [13] and Goldwave [14] were used to assist the extraction process, as well as the processes of down-sampling and up-sampling.

MFCCs have been shown to provide good comparison performance in the FVC arena [4, 15]. To calculate MFCCs, first the speech signal is converted into the frequency domain using the Discrete Fourier Transform (DFT). Then the energies in various regions of the frequency domain are estimated over a set of overlapped Mel-filter banks. MFCCs are then calculated by taking a Discrete Cosine Transform (DCT) of the logarithm of energies. Different numbers of MFCCs (typically, between 12 to 16) have been used by researchers [16, 17]. At a very early stage of this study, some FVC experiments were undertaken using different numbers of MFCCs. Based on the results of those experiments, it was decided to use 14 MFCCs in our token-based experiments. To take advantage of the delta-Cepstral feature, 14 deltas and 14 delta-deltas were added to the 14 MFCCs in the data-stream-based analysis.

3.2. *Comparison Process*

The 130 male speakers were divided into three mutually exclusive sets: 44 speakers for the Background set and 43 speakers each for the Development and Testing sets. (Note: the FVC results from the Development set are used to calibrate and fuse the results from the Testing set [9]). Data from three of the four recording sessions were used for the speakers in the Background set, while all four recording sessions were used for each of the speakers in the Development and Testing sets. The Suspect model for each comparison was formed using data from recording Sessions 1 and 2. This gives eight tokens per vowel for token-based analysis and eight utterance-segments per speaker for data-stream-based analysis. Sessions 2, 3 and 4 were used in turn for the Offender data. For the same-speaker comparisons, Sessions 3 and 4 for each speaker (i.e., Offender data) were compared with the Suspect model of the same speaker. For the different-speaker comparisons, Sessions 2, 3 and 4 for each speaker (i.e., Offender data) were compared with the Suspect

models for the other speakers. More details of these comparisons can be found in [18].

With 43 speakers in each of the Testing and Development sets, 43 same-speaker comparisons and 903 different-speaker comparisons are possible (ignoring multiple comparisons required in order to compute the CI). The results for individual vowels for the token-based analyses were then calibrated and fused, but for data-stream analyses, the results were just calibrated. Calibration and fusion were achieved using logistic regression [9]. The C_{llr} was calculated from the average of LRs for the two same-speaker comparisons and the average of LRs for the three different-speaker comparisons. The CI for both same-speaker and different-speaker comparison results were computed using the procedure outlined in [8].

3.3. *Experimental Set-up for Addressing f_s-Mismatch between Suspect, Offender, and Background Data*

To address f_s-mismatch between speech data sets in a FVC experiment, three scenarios have been investigated: (a) the Offender data sampled at 8 kHz, with the Suspect and Background data sampled at 32 kHz. The Suspect and Background data were then down-sampled to 8 kHz to match the Offender data, (b) the Offender data originally at 8 kHz and then up-sampled to match the Suspect and Background data at 32 kHz, and (c) the Offender, Suspect, and Background data originally at 8 kHz (having previously been down-sampled from 32 kHz) and then up-sampled back 32 kHz. It is acknowledged that this last scenario is not very realistic in the sense of why would one want to do that in a real forensic case, but this experiment was carried out for the sake of completeness to see whether it produced any unexpected results. As an example, the block diagram for the set-up of Scenario (b) is shown in Fig. 1.

4. Results

Table 1 shows the mean C_{llr} and 95% CI for Scenarios (a) – (c). It is clear that Scenario (b) results in very poor FVC performance, as can be seen from mean C_{llr} values being close to 1 for token-based analysis and relatively high for data-stream-based analysis.

Table 1: *Mean C_{llr} and CI for investigations that addressed f_s-mismatch between speech data sets in a FVC*

Sampling-Frequencies		Token-based	Data-stream-based
Scenario (a): Offender Speech: 8 kHz Suspect & Background Speech: 32 kHz to 8 kHz	Mean C_{llr}	**0.214**	**0.008**
	95% CI	2.245	1.069
Scenario (b): Offender Speech: 8 kHz to 32 kHz Suspect & Background Speech: 32 kHz	Mean C_{llr}	**0.903**	0.796
	95% CI	0.561	0.270
Scenario (c): Offender Speech: 8 kHz to 32 kHz Suspect & Background Speech: 8 kHz to 32 kHz	Mean C_{llr}	**0.282**	**0.021**
	95% CI	3.485	2.257

Comparing Scenarios (a) and (c), Scenario (c) has clearly resulted in a worse performance than Scenario (a) for both token-based and data-stream-based analyses. But the degradation in performance is certainly not as severe as that for Scenario (b).

So, the obvious question is: What is it about Scenario (b) that has resulted in such poor FVC performance? In Scenario (a), the Offender, Suspect, and Background data sets have all been down-sampled to 8 kHz. The resulting loss of spectral information in the frequency band 4-16 kHz for all three data sets would thus be the same. (Note: with sampled signals it is only possible to retain spectral information up to half the f_s.) But with Scenario (b), the Offender data was first down-sampled to 8 kHz, thereby losing spectral information in the band 4-16 kHz, then up-sampled back to 32 kHz. But even though it has been up-sampled back to 32 kHz, the spectral information it lost in the 4-16 kHz band as a result of down-sampling remains lost. However, the Suspect and Background data remained at 32 kHz and thus suffered no loss of spectral information in the 4-16 kHz band. So, one is no longer comparing 'like with like' with Scenario (b). Specifically, the up-sampled Offender data contained speech spectral energy only in the band 0-4 kHz, whereas the Suspect and Background data contained speech spectral energy in the band 0-16 kHz. We hypothesize that this is the reason for the very poor FVC performance of Scenario (b).

Returning again to a comparison of Scenarios (a) and (c), it is surprising that the FVC performance for Scenario (c) is somewhat worse than for Scenario (a). For Scenario (a), all three data sets were down-sampled to 8 kHz, so they then contained speech spectral energy in the band 0-4 kHz. Though in Scenario (c) all three data sets were first down sampled to 8 kHz and then up-sampled back to 32 kHz, they still only contained speech spectral energy in the band 0-4 kHz. So, in both scenarios one should be comparing 'like-with-like'. One expects this up-sampling process to be reasonably transparent in terms of not losing or degrading information. But the results here suggest that in the FVC arena this is clearly not the case.

A final observation can be gleaned from Table 1, namely that data-stream-based analysis outperforms token-based analysis in all scenarios. Given that the former utilises much more information, this is to be expected.

Figs. 2 and 3 show Tippett plots for Scenarios (a) to (c) for token-based and data-stream-based analysis, respectively. Considering first the plots of Figs. 2 and 3 for Scenario (b), it is clear that both same- and different-speaker comparisons have been similarly and adversely affected, and there is a large number of misclassifications for both. In addition, both the red line (representing different-speaker comparisons) and the blue line (representing same-speaker comparisons) are very close to the green line (i.e., LLR=0). This means that the LLR (or LR) is not providing much useful speaker-specific information, and this is true for token- and stream-based analysis.

Comparing now Scenarios (a) and (c) in Figs 2 and 3, it can be seen that for token-based analysis there is an increase in the number of same-speaker misclassifications for Scenario (c) compared to Scenario (a). For data-stream-based analysis, there are no same-speaker misclassifications for either scenario, but the performance of same-speaker comparisons for Scenario (a) is better than for Scenario (c). In terms of reliability, it can be seen that the 95% CI in Scenario (c) for both same- and different-speaker comparisons is larger than it is for Scenario (a).

5. Conclusions

This paper has addressed f_s-mismatch between Suspect, Offender, and Background data sets in a FVC and what strategy should be employed to correct for this. The LR

framework has been used for the experiments, and GMM-UBM has been used to compute LR values both for token-based and data-stream-base analyses.

As the XM2VTS, the speech database used in our study, contains speech data sampled at 32 kHz, this is the highest f_s used in our experiments. Speech data was then down-sampled to 8 kHz and then up-sampled again depending upon the scenario to be investigated. MFCCs were the speech features used in our investigations.

Three scenarios have been investigated: (a) the Offender data at 8 kHz and the Suspect and Background data down-sampled to 8 kHz to match, (b) the Offender data up-sampled from 8 kHz to 32 kHz to match the Suspect and Background data at their original 32 kHz, and (c) the Suspect, Offender, and Background data all up-sampled from 8 kHz to 32 kHz, having previously been down-sampled from 32 kHz to 8 kHz.

The results of our experiments conclusively show that Scenario (a) is the strategy to be adopted. The converse, i.e., Scenario (b), results in exceedingly poor FVC performance because, we hypothesise, one is then not comparing 'like-with-like'. Our results have also shown that up-sampling of all data sets from 8 kHz to 32 kHz, i.e., Scenario (c), results in poorer FVC performance, though not significantly, than leaving them all at 8 kHz. So, we conclude that there is no advantage in doing this.

The speech database used in this research is not very forensically realistic as it contains good quality speech recordings. When using more forensically-realistic data, we expect to get somewhat poorer results, but nonetheless, we do expect to observe the same trend. The reason is that our observed trend is related to a fundamental issue (i.e., spectral-data mismatch) and not the quality of the speech data.

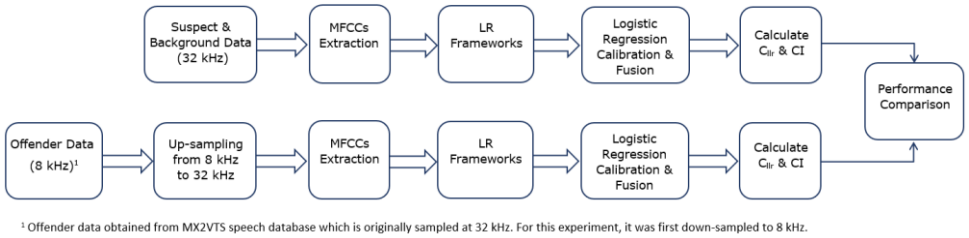


Figure 1: *Experimental set-up for addressing f_s-mismatch between speech data sets in a FVC (Scenario (b))*

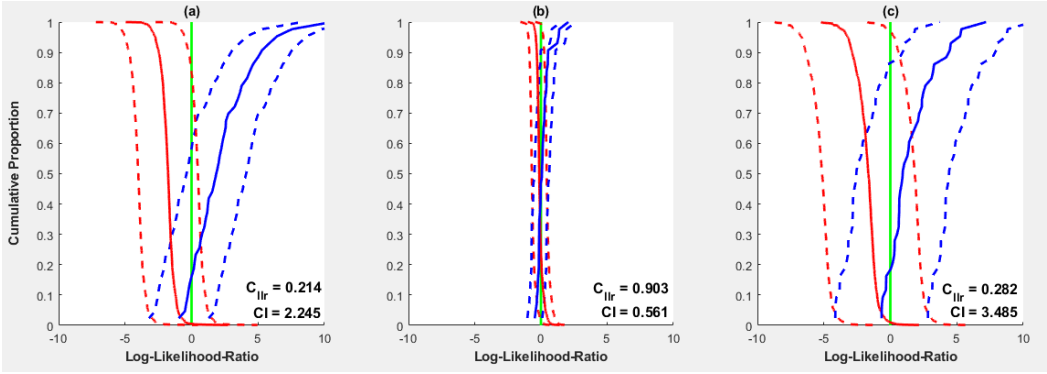


Figure 2: *Tippett plots showing the performance of GMM-UBM when applied to token-data: (a) Offender speech data at 8 kHz, Suspect and Background speech data are down-sampled from 32 kHz to 8 kHz; (b) Background and Suspect speech data are originally at 32 kHz, Offender speech data are up-sampled from 8 kHz to 32 kHz; (c) all three data sets are up-sampled from 8 kHz to 32 kHz*

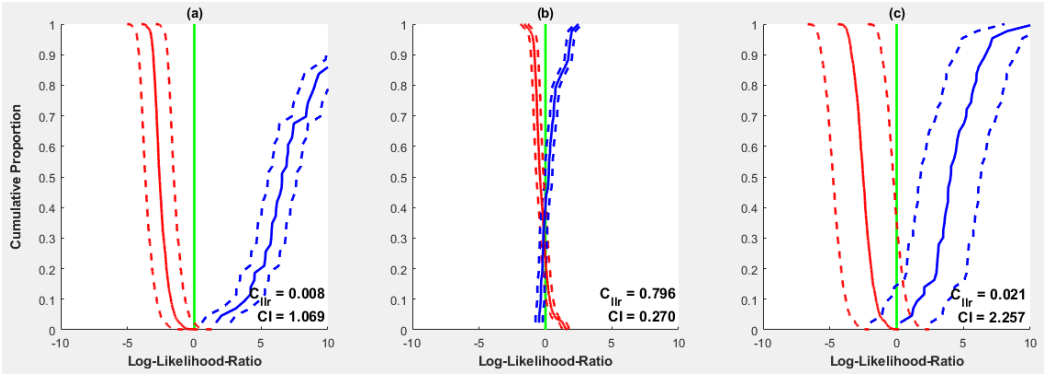


Figure 3: *Tippett plots showing the performance of GMM-UBM when applied to stream-data: (a) Offender speech data at 8 kHz, Suspect and Background speech data are down-sampled from 32 kHz to 8 kHz; (b) Background and Suspect speech data are originally at 32 kHz, Offender speech data are up-sampled from 8 kHz to 32 kHz; (c) all three data sets are up-sampled from 8 kHz to 32 kHz*

6. References

1. Reynolds, D., *Gaussian Mixture Models,* . Encyclopedia of Biometric Recognition, Springer, 2008.
2. Reynolds, D.A., T.F. Quatieri, and R.B. Dunn, *Speaker verification using adapted Gaussian mixture models.* Digital signal processing, 2000. **10**(1): p. 19-41.
3. Reynolds, D.A. *Automatic speaker recognition using Gaussian mixture speaker models*. in *The Lincoln Laboratory Journal*. 1995. Citeseer.
4. Morrison, G.S., *A comparison of procedures for the calculation of forensic likelihood ratios from acoustic–phonetic data: Multivariate kernel density (MVKD) versus Gaussian mixture model–universal background model (GMM–UBM).* Speech Communication, 2011. **53**(2): p. 242-256.
5. Dempster, A.P., N.M. Laird, and D.B. Rubin, *Maximum likelihood from incomplete data via the EM algorithm.* Journal of the Royal Statistical Society: Series B (Methodological), 1977. **39**(1): p. 1-22.
6. Hastie, T., R. Tibshirani, and J. Friedman, *The elements of statistical learning: data mining, inference, and prediction*. 2009: Springer Science & Business Media.
7. Morrison, G.S., *Forensic voice comparison and the paradigm shift.* Science & Justice, 2009. **49**(4): p. 298-308.
8. Morrison, G.S., *Measuring the validity and reliability of forensic likelihood-ratio systems.* Science & Justice, 2011. **51**(3): p. 91-98.
9. Morrison, G.S., *Tutorial on logistic-regression calibration and fusion: converting a score to a likelihood ratio.* Australian Journal of Forensic Sciences, 2013. **45**(2): p. 173-197.
10. Meuwly, D. and A. Drygajlo. *Forensic speaker recognition based on a Bayesian framework and Gaussian Mixture Modelling (GMM)*. in *2001: A Speaker Odyssey-The Speaker Recognition Workshop*. 2001.
11. Messer, K., et al. *XM2VTSDB: The extended M2VTS database*. in *Second international conference on audio and video-based biometric person authentication*. 1999. Citeseer.
12. Morrison, G.S., P. Rose, and C. Zhang, *Protocol for the collection of databases of recordings for forensic-voice-comparison research and practice.* Australian Journal of Forensic Sciences, 2012. **44**(2): p. 155-167.
13. Wavesurfer, *Wavesurfer. (2012, December 5 – last update). Retrieved on 12 January 2013, last retrieved from http://www.speech.kth.se/wavesurfer/index2.html*. 2011.
14. GoldWave, *GoldWave. (2013, January 24 – last update). Retrieved on 5 March 2013, last retrieved from http://www.goldwave.com/*. 2012.
15. Nair, B.B., E.A. Alzqhoul, and B.J. Guillemin. *Comparison between Mel-frequency and complex cepstral coefficients for forensic voice comparison using a likelihood ratio framework*. in *Proceedings of the World Congress on Engineering and Computer Science, San Francisco, USA*. 2014.
16. Rabiner, L.R. and R.W. Schafer, *Introduction to digital speech processing*. 2007: Now Publishers Inc.
17. Vandyke, D.J., *Glottal Waveforms for Speaker Inference & A Regression Score Post-Processing Method Applicable to General Classification Problems*. 2014, Citeseer.
18. Mehdinezhad, H. and B.J. Guillemin, *Preliminary performance comparison between PCAKLR and GMM-UBM for computing the strength of speech evidence in forensic voice comparison.* SST2016, 16th Speech and Science Technology, 2016.

Tone and Vowel Interaction in Northern Lisu

Rael Stanley, Marija Tabain, Defen Yu, David Bradley

Department of Languages and Linguistics, La Trobe University, Melbourne, Australia

r.stanley@latrobe.edu.au, m.tabain@latrobe.edu.au, d.yu@latrobe.edu.au, d.bradley@latrobe.edu.au

Abstract

This study presents data for f0 and H1*-H2* in Northern Lisu Tones, and compares the results between back and front vowels. F0 and H1*-H2* tracks are extracted for four of the six tones in Northern Lisu - 33, 33̲, 21, 21̲ʔ (where underlining indicates creaky voice quality). These are compared between tones that appear on raised vowels /ɯ, u, ɤ, o/, front vowels /i, y, e, ø, ɛ/, and the retracted vowel /ɑ/. Results show that there is a difference in tone contour when comparing across the three different vowel contexts.

Index Terms: acoustic phonetics, tone, voice quality, Northern Lisu, Tibeto-Burman languages, vowels

1. Introduction

Lisu is a language in the Loloish branch of the Tibeto-Burman language family [1] and has speakers in Northeast India, Northern Myanmar, Southwest China, and Thailand [2]. There are three main dialects of Lisu - Northern, Central, and Southern [3].

Most varieties of Lisu have six tones, though some have fewer [2]. Northern Lisu is one of the varieties with six tones, and they are: high-level (55); mid-rising (35); mid-level (33); mid-level, creaky voiced (33̲); low-falling (21); and low-falling, creaky voiced with final glottal stop(21̲ʔ) [2]. However, speakers of Northern Lisu have a tendency to merge the 33 and 33̲ tones [3].

There are 10 monophthongal vowels in Northern Lisu /i, y, e, ø, ɛ, ɑ, ɤ, o, ɯ, u/, as well as a fricative vowel /ʐ̩/ [2]. However, the contrast between /ɯ/ and /ɤ/ is a marginal one. There are also diphthongs /iɑ/ and /uɑ/. However, these do not appear in native Lisu words and occur only in loanwords [4, 2, 3].

In Esling et. al.'s Laryngeal Articulator Model [5], the authors classify vowels into categories which account for both lingual and laryngeal properties. Using this model, the vowels of Northern Lisu fall into three categories: Front vowels /i, y, e, ø, ɛ/, Raised vowels /ɯ, u, ɤ, o/, and the Retracted vowel /ɑ/. According to this model, retracted vowels may be more likely to be affected by laryngeal constriction, resulting in a higher likelihood of creaky voice, we explore this possible interaction with a language that has a relatively large set of tonal contrasts, as well as front, raised, and retracted vowel contrasts.

2. Method

2.1. Speakers

All participants were native speakers of Northern Lisu, aged between 20 and 22 at the time of recording in 2017 (N=8, four female speakers and four male speakers). They were all born in the Nujiang Lisu Autonomous Prefecture, in Western/Northwestern Yunnan Province, China, and all were students at Yunnan Minzu University, except for one who was a student at Yunnan Normal University, both located in the Chenggong district of Kunming.

2.2. Word List

Participants were recorded reading from a list of target words in Northern Lisu that was generated using [4]. For this analysis, each word in the list was selected in such a way that each of Northern Lisu's ten monophthongal vowels (excluding the fricative vowel) was accompanied by each of the six tones. In order to minimise co-articulatory effects, every effort was made to choose words where a bilabial stop preceded the target vowels. However, this was not possible in every case. In total, there were 54 words in the list.

The words were then translated into Chinese orthography, so that any influence from Lisu orthography could be minimised.

Participants were asked to say each word three times in isolation, in order to avoid potential effects of tone sandhi. However, not all participants knew every word, and some words had multiple potential Northern Lisu words to choose from. Because of this, not every speaker produced tokens of every word listed, and some words produced were different to what may have been expected based on the entry in the dictionary. In these cases, these unexpected words were noted and compared with the Sino-Tibetan Etymological Dictionary and Thesaurus (STEDT) [6]. Where the production clearly matched one of the various definitions for the same word in the STEDT, it was included for analysis. In cases where the produced word did not match an entry in the STEDT, it was left out of the analysis, so that judgements on tone or vowel quality by a non-native speaker would not influence the results. In total there were 969 tokens analysed.

2.3. Recordings

Recordings were made using an H4n Zoom in a quiet hotel room in Kunming. All recordings were done in mono with a sampling rate of 48 kHz (with one exception, due to the Zoom running low on battery and automatically lowering the sampling rate to 24 kHz, which was later upsampled to 48 kHz using Audacity, so that the affected speaker could be included with other speakers in the analysis), and a bit depth of 1536 kbps. These files were saved as uncompressed .wav files.

2.4. Analyses

Prior to analysis, each participant's recording was split into individual utterances using Praat [7], by utilising scripts to detect

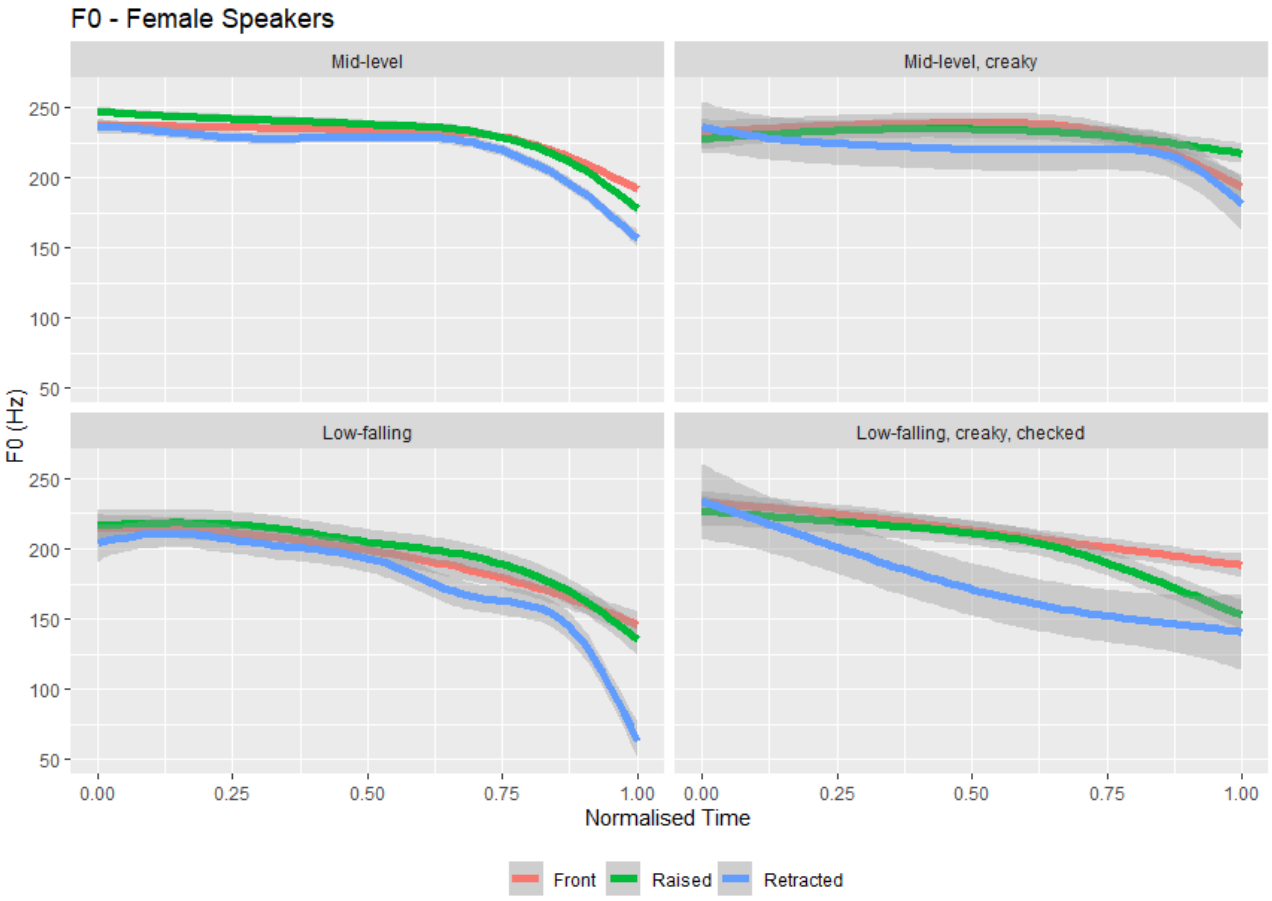


Figure 1: F0 across normalised time for female speakers of Northern Lisu

pauses, mark boundaries, and split the file into individual words [8]. TextGrids were then created using Praat, with tiers: Word, Sex, Tone, Repetition, Segment, Speaker, and Syllable.

Pitch traces across normalised time were produced using the emuR package [9] in R [10]. H1*-H2* values were produced using VoiceSauce [11] in MATLAB [12]. Plots were made using the ggplot2 [13] package. F0 and H1*-H2* data will be presented as GAM (Generalised Additive Model)-smoothed trajectories, with grey intervals indicating 95% confidence levels.

3. Results

3.1. Female Speakers - F0

As can be seen in the top panels of Figure 1, there is little difference in f0 contour for mid (top-left panel) versus mid-level creaky (top-right panel) tones, of female speakers. In both these tones, f0 values in the retracted vowel context (blue line) are overall lower than the front (red line) or raised (green line) vowel contexts. For the mid-level tone, f0 values in the retracted vowel context are significantly lower than in either of the raised or front vowel contexts from approximately 70% of total duration, until the end of the vowel.

In the low-falling (bottom-left panel) and low-falling creaky checked (bottom-right panel) tones, f0 values are also overall lower in the retracted vowel context than in the other vowel contexts. In the low-falling tone, there is a signifcant drifference from approximately 70% of total duration until the end of the vowel, as with the mid-level tone.

The low-falling creaky checked tone has a more complicated difference in these vowel contexts than the other tones. Between approximately 35% and 95% of total duration, f0 in the retracted vowel context is significantly lower than in the raised or front vowel contexts. From approximately 80% of total duration until the end of the vowel, the low-falling creaky checked tone has a significantly higher f0 in the front vowel context than either of the other vowel contexts.

3.2. Male Speakers - F0

As can be seen in Figure 2, male speakers produce all four of the tones examined with an overall lower f0 in the retracted vowel context as opposed to the front or raised vowel contexts, like the female speakers.

For the mid-level tone (top-left panel), productions in the retracted vowel context are significantly lower than in the front or raised vowel contexts, from the beginning of the vowel until approximately 75% of total duration, after which point there is no difference between the retracted vowel context and the raised vowel context. From approximately 70% of total duration until the end of the vowel, the mid-level tone has a significantly higher f0 in the front vowel context than in the other vowel contexts.

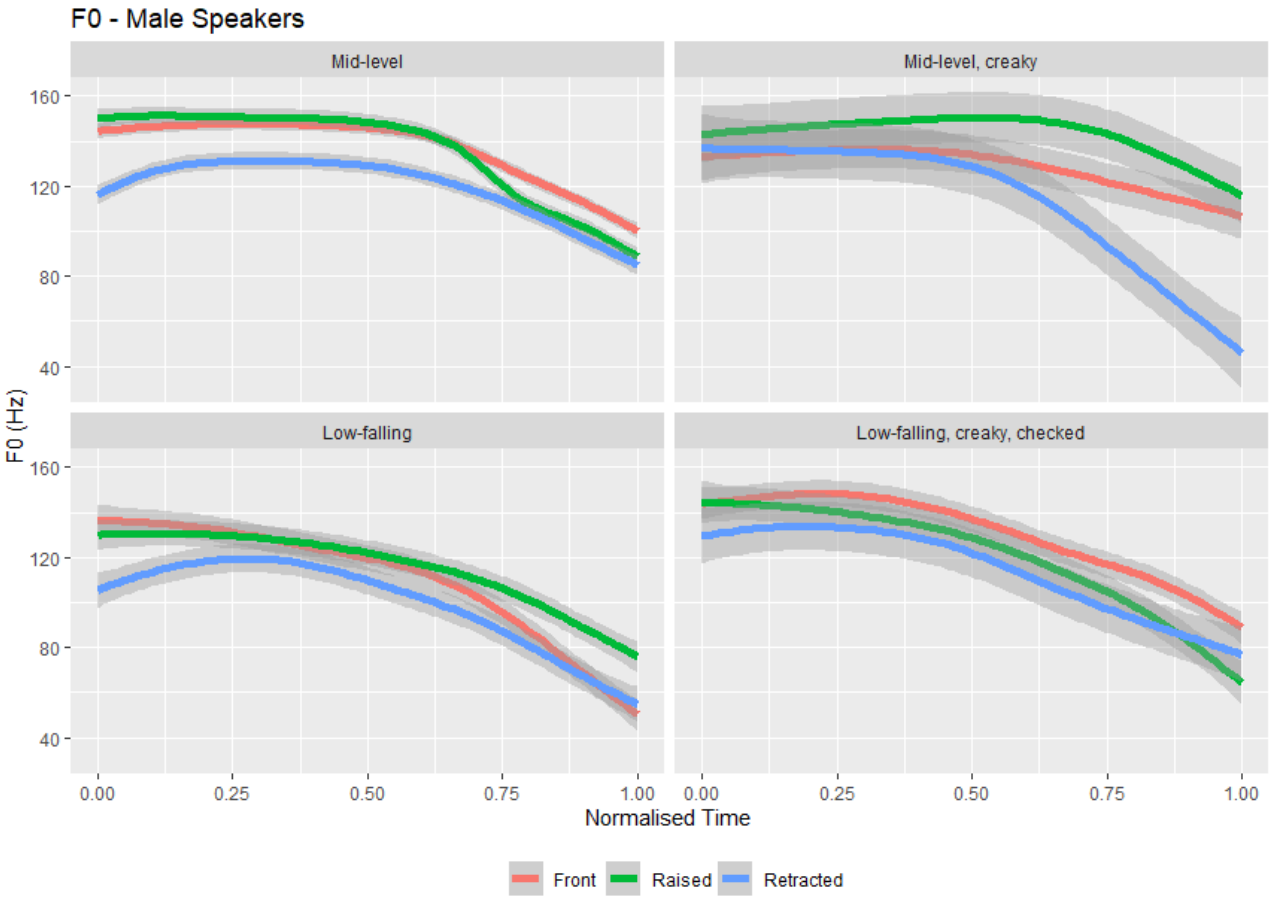


Figure 2: F0 across normalised time for male speakers of Northern Lisu

For the mid-level creaky tone (top-right panel) f0 of productions in the retracted vowel context are significantly lower than in the other vowel contexts from approximately 70% of total duration, until the end of the vowel.

For the low-falling tone (bottom-left panel), f0 of in the retracted vowel context is significantly lower for the first 20% of total duration, and for the last 20% of duration f0 is significantly higher in the front vowel context than in the other vowel contexts.

While f0 of the low-falling creaky checked tone is overall lower in the retracted vowel context, it is only significantly different when compared to the front vowel context and not the raised vowel context.

3.3. Male Speakers - H1*-H2*

As can be seen in Figure 3, H1*-H2* values for all tones except for the low-falling creaky checked tone are overall lower in the retracted vowel context than in the front or raised vowel contexts.

For the mid-level tone (top-left panel), H1*-H2* values are significantly lower in the retracted vowel context for almost the entire duration of the vowel. There is also a short portion of the vowel for which H1*-H2* values are significantly higher in the raised vowel context than in the front vowel context, from approximately 5% to 15% of total duration.

For the mid-level creaky tone (top-right panel), there is no significant difference in H1*-H2* values between productions in the retracted vowel context and the front vowel context. However, H1*-H2* values for both these vowel contexts are significantly lower than the raised vowel context from approximately 25% of total duration until the end of the vowel (aside from a small degree of overlap in the grey confidence intervals between the front and raised vowel contexts between 55% and 80% of total duration).

For the low-falling tone (bottom-left panel), H1*-H2* values in the retracted vowel context are significantly lower than the other two vowel contexts from the start of the vowel until approximately 20% of total duration. After this point, there is no significant difference between the retracted and front vowel contexts. However, both these vowel contexts are significantly lower in H1*-H2* values than the raised vowel context from approximately 15% of total duration until the end of the vowel.

For the low-falling creaky checked tone (bottom-right panel), there is no difference in H1*-H2* values between the retracted and raised vowel contexts. However, between approximately 15% and 50% of total duration H1*-H2* values are higher in the front vowel context than in either the raised ro retracted vowel contexts.

Due to reasons of space, only male data is presented here. The female plots show differences based on vowel context. However, it appears that tones in the retracted vowel context tend to be breathier (that is, displaying an overall higher H1*-

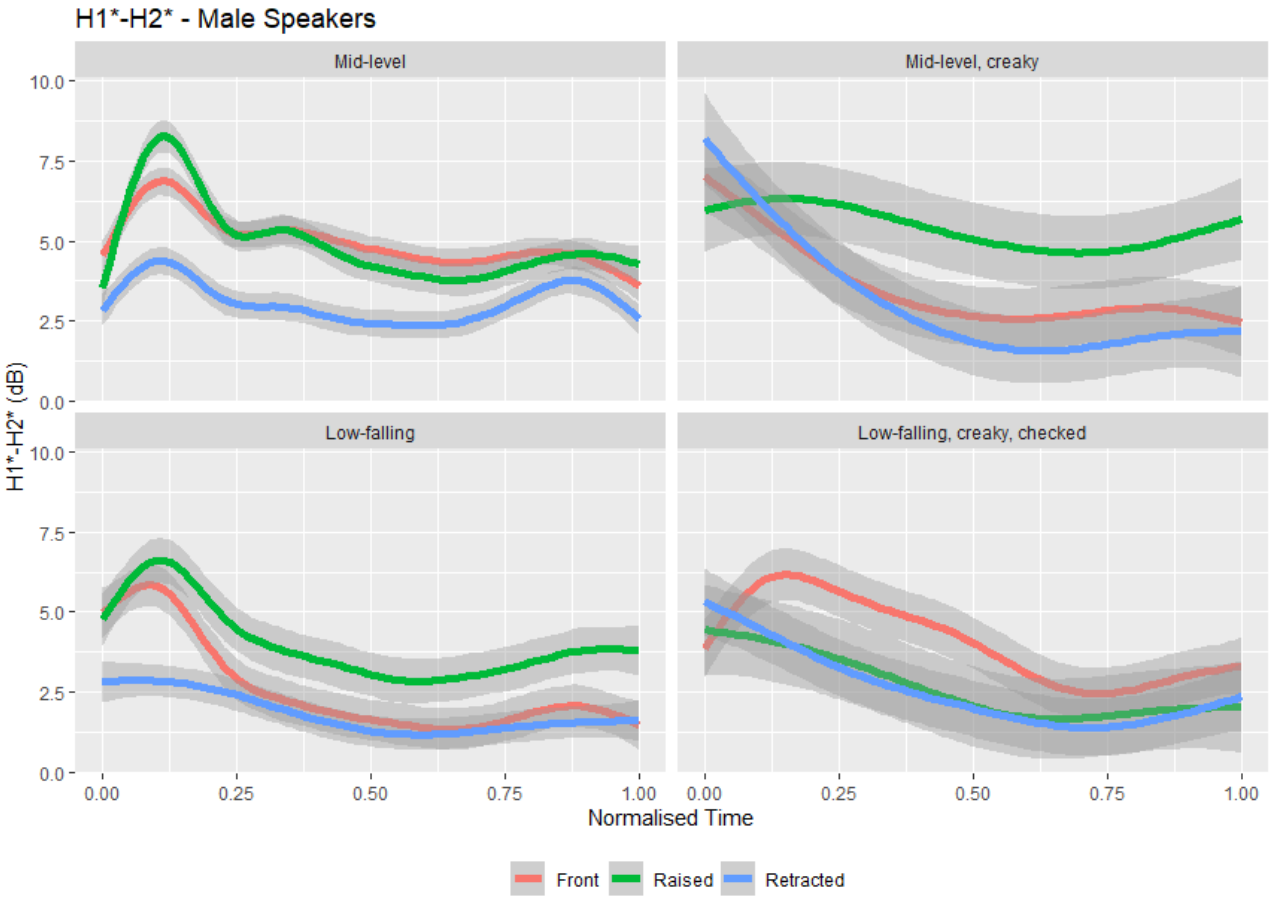


Figure 3: H1*-H2* across normalised time for male speakers of Northern Lisu

H2*) than in the other vowel contexts.

4. Discussion

As shown above, vowel context affects f0 in both the mid-level tone pairs and the low-falling tone pairs, in Northern Lisu, resulting in lower overall f0 for both in the retracted vowel context. It has a greater effect on the f0 of low-falling, creaky, checked tones, making them lower overall than the low-falling tones, for female speakers. For male speakers, this effect is stronger in the mid-level and mid-level creaky tones.

Female speakers displayed almost no difference between the mid-level and the mid-level, creaky tones, overall. This indicates that for female speakers, there is further evidence of merger between these two tone categories.

Male speakers, however display larger differences between mid-level and mid-level, creaky tones. This may be another example where female speakers appear to be leading a sound change in progress (i.e. the female speakers appear to be merging the mid-level and mid-level creaky tones, as had previously been noted by Bradley [3]).

These results also fit with Esling et. al.'s [5] conclusion that the laryngeal constriction involved in producing retracted vowels shortens vocal folds and lowers pitch. There is also some evidence for this constriction influencing creakiness, with H1*-H2* values in this vowel context being overall lower in even the modal tones. Although, the nature of this interaction is, as Esling states, a complex one.

5. Acknowledgements

We would like to thank He Jian Xin and Hu Shi Yu for their patience and assistance in the data recording, as well as the participants for their hard work.

6. References

[1] Bradley, D., "Lisu", in G. Thurgood and R. LaPolla [Eds], The Sino-Tibetan Languages, 222-235, Routledge, 2003.

[2] Yu, D., Aspects of Lisu Phonology and Grammar, La Trobe, 2007.

[3] Bradley, D., "Lisu Language", in R. Sybesma, W. Behr, Y. Gu, Z. Handel, and C. -T. Huang [Eds], Encyclopedia of Chinese Languages and Linguistics, Brill, 2018.

[4] Bradley, D., A Dictionary of the Northern Dialect of Lisu (China and Southeast Asia), Australian National University, 1994.

[5] Esling, J.H., Moisik, S.R., Benner, A., and Crevier-Buchman, L., The Laryngeal Articulator Model, Cambridge University Press, 2019.

[6] STEDT Project, "Sino-Tibetan Etymological Dictionary and Thesaurus", Online: https://stedt.berkeley.edu, accessed on 4 May 2022.

[7] Boersma, P. and Weenink, D., Praat: doing phonetics by computer [Computer program]., Verstion 6.2.12, Online: http://www.praat.org/, retrieved 17 April 2022.

[8] Lennes, M., SpeCT - Speech Corpus Toolkit for Praat (v1.0.0), Online: https://doi.org/10.5281/zenodo.375923, accessed 4 May 2022.

[9] Winkelmann, R., Jaensch, J., Cassidy, S., and Harrington, J., emuR: Main Package of the EMU Speech Database Management System., R package version 2.3.0, 2021.

[10] R Core Team, R: A Language and Environment for Statistical Computing., Online: https://www.R-project.org., accessed 4 May 2022.

[11] Shue, Y.-L., Keating, P., Vicenik, C., and Yu, K., VoiceSauce: A Program for Voice Analysis, Proceedings of the ICPhS XVII, 1846-1849, 2011.

[12] MATLAB., version 7.10.0 (R2010a), Natick, Massachusetts: The MathWorks Inc., 2010.

[13] Wickham, H., ggplot2: Elegant Graphics for Data Analysis., Springer-Verlag., Online: https://ggplot2.tidyverse.org., accessed 4 May 2022.

[14] Pinheiro, J.C. and Bates, D.M., Mixed-Effects Models in S and S-Plus., Springer, doi: 10.1007/b98882., accessed 4 May 2022.

Tonal effects on vowel duration in Bangkok Thai

Francesco Burroni[a*], *Teerawee Sukanchanon*[b*]

[a]Department of Linguistics and Cognitive Science Program, Cornell University, USA
[b]Southeast Asian Linguistics Research Unit, Faculty of Arts, Chulalongkorn University, Thailand
fb279@cornell.edu, t.sukanchanon@gmail.com

Abstract

We investigated tonal effects on vowel duration in two experiments with 37 speakers of Bangkok Thai. For long vowels in open syllables, the pattern of tonal effects on vowel duration is {Mid,Low} < {Falling,High,Rising}; for closed syllables with short vowels and sonorant codas, the pattern is {Rising} < {Mid,Low} < {High} < {Falling}. Our results do not align with hypothesized universal patterns or with previous reports on Bangkok Thai. Our findings are better understood by referring to f0 control mechanisms. Finally, we found that the tonal effects are mediated by syllable structure in line with diachronic changes in vowel length.

Index Terms: tone, vowel duration, vowel length, word duration, fundamental frequency, Thai

1. Introduction

A claim often repeated in the literature,e.g., [1], and going back to [2], is that the effects of tone on vowel duration are hypothesized to be universally inversely proportional to average fundamental frequency (f0) values. Vowels in syllables with higher f0 have shorter vowel duration, while those in syllables with lower f0 have longer duration. Gandour [2] based his account on diachronic data from Thai varieties, where long vowels emerged in association with rising and non-high level tones, and short vowels emerged in association with falling and high level tones. Gandour's claim was also in line with Abramson's findings on Bangkok Thai (BKKT) [3], where the Mid (M) and Low (L) tones, tones with relatively low average f0, are longer than the Falling (F) and High (H) tones, tones with relatively high average f0. Note, however, that none of the BKKT tone is level, hence, the relationship between their names and f0 contour should be taken with a grain of salt, Figure 1.

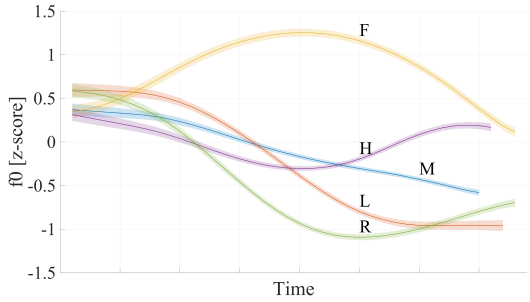


Figure 1: *Average BKKT tonal contours from exp. 1*

Subsequent work showed that Gandour's hypothesis may not hold for all languages. One study has shown that among the three level tones in Taiwanese, the durational effect hierarchy is L < M < H [4]. One study of Cantonese [5] has shown that the duration effect hierarchy is {L,H} < ML < M (braces are used to indicate sets of tones that do not differ). There are reasons to question whether Gandour's hypothesis and the order of durational effects reported in [3] are correct, even for BKKT. The first issue is based on physiological considerations. The durational effects of tone reported by [3], {F,H} < {M,L}, are the opposite of what f0 control mechanisms would predict; as it has been reported that f0 falls are faster/shorter than f0 rises [6], [7]. If the effects of tone on duration are related to f0 control, the M and L tones, which only have a falling component, should be associated with shorter vowel durations. A second issue comes from diachronic data. In BKKT, the words that underwent vowel lengthening are almost only words with tones that are relatively high in the f0 range, the F and H tones [8]. What emerges from diachronic data is also that all lengthened words have /a/ as nucleus and sonorants as codas, for the most part, /-j/ and /-w/. BKKT vowels also underwent shortening. Syllables ending with nasals /-m, -n, -ŋ/ were typically shortened when occurring with the L and F tones [8]. Tone-coda interaction effects suggest that the durational effects of tone may be conditioned by phonological factors, like syllable structure and coda types. A third issue that has not been explored is whether durational effects of tone may similarly affect the short and long vowels of BKKT, given that the language has a phonological contrast for vowel length.

Against this background, we formulate the following research questions: (I) Is the order reported by [3] for tonal effects on vowel duration still observed in contemporary BKKT? (II) What is the relationship, if any, between the durational effects of tone, f0 control mechanisms, and diachronic changes in vowel length? (III) Are the effects mediated by other phonological factors, e.g., vowel length and syllable structure? We make the following hypotheses and associated predictions. For (I), if the hypothesis of [2], an inverse relationship between tonal durational effects and average f0, is correct; then, we expect to observe the durational effect hierarchy reported in [3]. The M, L, and R tones, which have relatively low average f0, should be associated with longer vowels. However, the diachronic changes in vowel length and known differences between the speed of f0 rises and falls may suggest otherwise. For (II), if the order of tonal duration effects reported in [3] is correct, we do not have any clear relationship between f0 control mechanisms and vowel duration. We also do not have any clear association with diachronic changes. If, however, f0 control mechanisms are one of the factors determining vowel duration and, moreover, diachronic changes reflect synchronic phonetic variation, we may expect an order of tonal durational effects like {M,L,F} < {H,R} (f0 falls are shorter and f0 rises are longer [7]) or {R} < {M,L} < {F,H} (R induces shortening, while F and H induce lengthening, as diachronic data show). For (III), the diachronic data suggest that

[*] Equal contribution.

certain combinations of tones and codas result in phonological lengthening or shortening diachronically. We may, thus, expect to observe traces of these interactions in synchronic variation in vowel duration. We have investigated these questions by studying the vowel /a/, the only vowel in BKKT that has undergone phonological lengthening diachronically. We have studied two environments: /Ca:/ (long /a/ in open syllables) and /Ca(G|N)/ (short /a/ in closed syllables with glide or nasal codas); as these are among the only syllabic configurations where all 5 BKKT tones (M, L, F, H, R) can be licensed.

2. Methods

For experiment 1 (exp. 1), data were collected from 20 speakers (range=19-52, mean=28.4, std=12.8), and, for experiment 2 (exp. 2), from 17 speakers (range=20-23, mean=21.2, std=0.8). Participants in exp. 1 were students or staff at a North American university; participants in exp. 2 were students at a university in Bangkok. All speakers were screened for nativeness in BKKT by a native speaker trained in phonetics. They reported no speech or hearing impairments. In exp. 1, participants produced [ma:] with all five tones of BKKT (M, L, F, H, R). The target was embedded in a carrier sentence with a fixed number of words and syllables, Table 1.

Table 1. *Stimuli in exp. 1*

w1	w2	w3	w4	w5
dū:	mî:	mā:	bōn	bɔ̄n
	mǐ:	mà:		dā:w
		mâ:		lāŋ
		má:		
		mă:		

Disyllabic tonal combination of w2 and w3 are intended to represent nonce words. F/R in w2 were chosen because the dataset was originally designed to study tonal coarticulation in separate work by one of the authors. Participants were instructed so that the F/R initial word of the combination represents an imaginary animal while the second one a fur pattern. Thus, the disyllabic tonal combinations represent noun-noun compounds, meaning "(I) look at a mî:/mǐ: (with fur pattern) mā:/mà:/mâ:/má:/mă:". Participants completed 10 blocks in which they produced random combinations of 2 (F/R) × 5 (M/L/F/H/R) × 3 (distractors) = 30 unique stimuli. w5 of the carrier sentence was varied at every trial to function as a distractor. The number of tokens collected in exp. 1 was 20 (participants) × 30 (unique stimuli combinations) × 10 (blocks) = 6000. In exp. 2, the targets were 45 unique monosyllabic words with syllable structure /Caŋ/ and /CaG/ (where C = /p, t, k/ and G = /j, w/) combined with all 5 tones of BKKT. The word list was designed to maximize the number of nonce words. An example of all targets with /p/ onset is given in Table 2. Identical combinations were elicited with /k/ and /t/ onsets, yielding 3 (onsets) × 3 (rhymes) × 5 (tones) = 45 unique targets.

Table 2. *Example stimuli with p onset in exp. 2*

syllable structure		tones				
onset	rhyme	1	2	3	4	5
p	-aŋ	pāŋ	pàŋ	pâŋ	páŋ	păŋ
	-aj	pāj	pàj	pâj	páj	păj
	-aw	pāw	pàw	pâw	páw	păw

The targets were embedded in a carrier sentence [dū: khām wâ: X bōn Y] "Look at the word X on Y", in which X is the target and Y is one of eight disyllabic meaningful words acting as a distractor. The distractors are [pʰē:.dā:n] 'ceiling', [kām.phē:ŋ] 'wall', [thā:ŋ.dɤ̄:n] 'path', [kā:ŋ.kē:ŋ] 'trousers', [lām.pʰō:ŋ] 'speaker', [bān.dāj] 'stairs', [sāʔ.pʰā:n] 'bridge', and [kāʔ.lā:] 'coconut shell'. Exp. 2 was divided into 8 blocks. Each block contained all unique targets, pseudo-randomized, with a constraint that targets with the same tone will not appear consecutively. The distractors were evenly distributed across all tokens. The number of collected tokens in exp. 2 was 17 (participants) × 45 (unique stimuli) × 8 (blocks) = 6120.

For both experiments, participants sat in a sound-attenuated room in front of a computer monitor. A custom MATLAB GUI for each experiment was used to present stimuli in the form of pictures and phrases in Thai orthography. Simultaneously audio was collected, using a head-mounted microphone at a sampling rate of 44.1 kHz and 24 bits per sample. Speaker-specific monophone HMMs were trained in Kaldi [9] and used to perform forced alignment separately by speaker. The training data were hand-segmented in PRAAT [10] using the waveform, spectrogram, and changes in intensity/formant trajectories. Boundaries between vowels and glides were identified as the midpoint of the first formant transition. Boundaries between vowels and nasals were based on the appearance of antiformants and spectral/amplitude changes. Following forced alignment, a proportion of segmented data was randomly selected to be visually inspected and hand corrected. The manually checked alignments were used to retrain HMMs and realign all the trials. This process was repeated until the automatically segmented data were comparable to manual annotation. Trials that contained disfluencies were excluded. The word [kâw] in exp. 2 was excluded, as participants consistently produced it with a phonologically long vowel. In total, we analyzed 5977 tokens from exp. 1 and 5956 tokens from exp. 2. The main dependent variable we analyzed is vowel duration of the targets as defined by segmental boundaries. For exp. 1, given the identical segmental composition of all targets, we also analyzed word duration. For exp. 1, linear mixed effect regressions were conducted to assess the effect of tone, a categorical variable (with reference set as the M tone), on vowel duration. Random intercepts and slopes by speakers were also included. For exp. 2, we had two categorical variables, tone (reference M tone) and coda type (reference /-ŋ/), as well as their interaction. For random effects in exp. 2, we compared various structures and found speakers and onsets intercepts and slopes to be the maximal structure justified by a loglikelihood ratio test. The final statistical model for each dependent variable was selected in a pruning stepwise procedure, that is, we compared nested models and removed one fixed effect at a time and checked for significance *via* loglikelihood ratio tests.

3. Results

In exp. 1, we found that tone has a significant effect on long vowel duration in BKKT ($\chi^2_{(4)} = 22.7$, $p < .0001$). The M and L tones are systematically associated with shorter vowels than the F, H, and R tones. No significant differences were observed between M *vs.* L or among F *vs.* H *vs.* R. The differences between {M,L} and {F,H,R} range between –2 and –20 ms. Since the average vowel duration for the L and M tones is estimated at ~224 ms, the difference in duration associated with the presence of {F,H,R} tones represents an increase in duration in the range of 1-10% of the vowel. A full pairwise comparison with 95% confidence intervals (CI) for the differences in vowel duration among all tones is presented in Table 3. Each cell represents the difference between the tone in each row minus the tone in each column. Longer duration of vowels associated

with {F,H,R} *vs.* {M,L} is evident by inspecting the distributions of all tokens in boxplots, Figure 2.

Table 3. *V: duration differences between each tone*

	M	L	F	H	R
M		*ns*	–6, –19	–4, –21	–2, –18
L	*ns*		–6, –18.5	–5.5, –20	–4, –17
F	6, 19	6, 18.5		*ns*	*ns*
H	4, 21	5.5, 20	*ns*		*ns*
R	2, 18	4, 17	*ns*	*ns*	

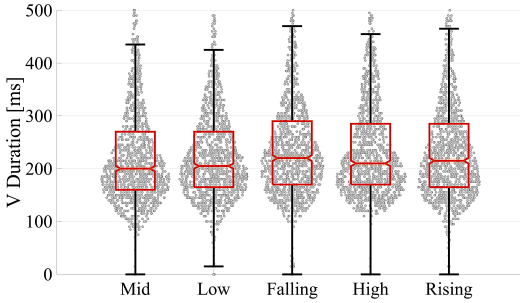


Figure 2: *V: duration in all tone conditions*

Word duration is also affected by tone category ($\chi^2_{(4)} = 23.6$, $p < .0001$) in the same way as vowel duration. M *vs.* L toned words are not different from each other. Nor are F *vs.* H *vs.* R toned words. However, {M,L} toned words are shorter than {F,H,R} toned words. 95% CI range between 5-25 ms; corresponding to an increase in duration for {F,H,R} toned words in the range of 2% to 8% of total word duration, estimated at ~314 ms.

In exp. 2, a significant effect of tone-coda interactions on vowel duration was observed ($\chi^2_{(10)} = 945.1$, $p < .0001$). Like in exp. 1, {M,L} are associated with shorter vowel duration than {F,H}. There are no differences between M *vs.* L or between F *vs.* H. However, we found that R toned vowels are significantly shorter than vowels associated with all other tones. Table 4 shows a full pairwise comparison with 95% CI for the differences in vowel duration among all tones (beige color indicates marginally significant differences).

Table 4. *V duration differences between each tone*

	M	L	F	H	R
M		*ns*	-2, -11	*0, -6*	1, 9
L	*ns*		-3, -13.5	-1, -8	0, 6
F	2, 11	3, 13.5		*ns*	6, 18
H	*0, 6*	1, 8	*ns*		4, 12
R	-1, -9	-0, -6	-6, -18	-4, -12	

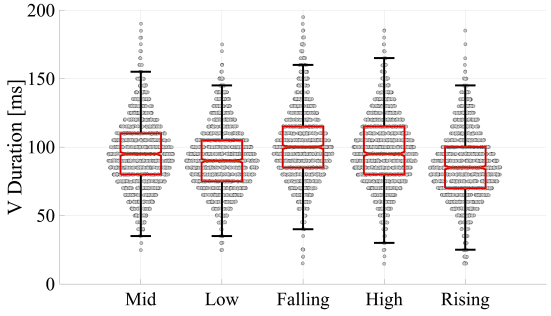


Figure 3: *V duration in all tone conditions*

For codas, vowel duration in syllables with bilabial glide coda (/-w/) is longer than in syllables with palatal glide (/-j/) and velar nasal (/-ŋ/) codas. Among the three codas, vowel duration in syllables with /-j/ is the shortest, Table 5 and Figure 4.

Table 5. *V duration differences between each coda*

	/-ŋ/	/-j/	/-w/
/-ŋ/		1, 6	-16, -21
/-j/	-1, -6		-20, -25
/-w/	16, 21	20, 25	

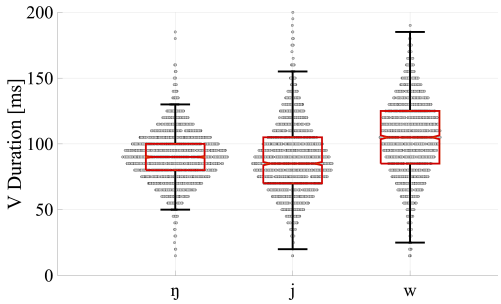


Figure 4: *V duration in all coda conditions*

We also observed coda and tone interaction effects on vowel duration. Significant differences in vowel duration due to tone-coda interactions along with 95% CI are presented in Table 6.

Table 6. *significant differences in vowel duration due to interactions (baseline is M tone with /-ŋ/ coda)*

	/-j/	/-w/
L	*ns*	-1, -8
F	7, 14	-3, -11
H	*ns*	*ns*
R	*ns*	-7, -14

The interactions can be interpreted as follows. /-j/ plus F is longer by [7, 14] ms than it is expected by adding the effects of /-j/ and F together. Notice how /-j/ coda plus F "stands out" more from the other /-j/ realizations even though the mean of all /-j/ realizations is lower than the other two coda conditions, orange arrows in Figure 5. A reverse effect is observed with /-w/, for which cooccurrence with the F, H, or R tones does not result in the degree of lengthening expected by adding the effects of coda and tone. In all these conditions the predicted vowel durations are shorter by [-8, -1] ms, [-11, -3] ms, and [-7, -14] ms respectively than predicted by a purely additive model. Note how /-w/ plus L/R are much further away from /-w/ plus M/H compared to their distances in the other two coda conditions, blue arrows in Figure 5. Finallly, note that /-w/ plus F does not "stand out" as much as expected from /-w/ with other tones, rightmost orange arrow in Figure 5.

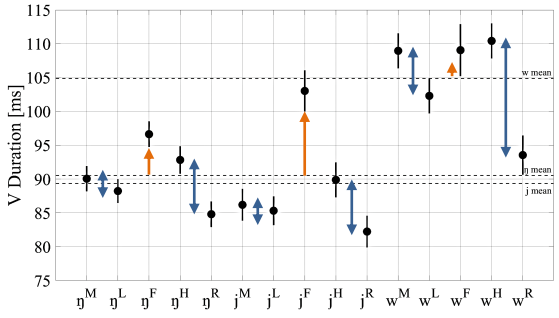


Figure 5: *Mean and 95% CI for all tones × codas*

4. Discussion

In exp. 1, we found that {M,L} toned long vowels in open syllables are systematically shorter than those in {F,H,R} toned syllables. In our data, the relative order of durational effects of tone on long vowels, {M,L} < {F,H,R}, is the *reverse* of what was reported by Abramson for BKKT [3]. These effects of tone on vowel durations are also problematic for the idea that such effects are universally inversely proportional to average f0 height of the tone [2]. In this respect, our data on long vowels in open syllables show that two tones with relatively low average f0 height, M and L, are linked to the shortest vowels. A more promising way to interpret our results is in terms of a transparent relationship between the durational effects of tone on vowels and f0 control mechanisms. It has been reported that, at least for untrained speakers, i.e., non-singers, f0 falls are produced faster than f0 rises [6], [7]. If the production of vowels and tones is temporally coordinated, so that their executions are time-locked or closely timed, our durational findings can be understood as follows. Vowels associated with the M and L tones, which comprise only faster f0 falling components are shorter; as these tones are produced with active f0 lowering laryngeal activity [11], [12]. While vowels with the F, H, and R tones, which all comprise slower f0 rising components either in the first half (F) or the second half (H, R) of the contour, Figure 1, are longer. Additionally, our findings are also in line with diachronic BKKT evidence. Vowel lengthening in BKKT is sporadic but observed almost exclusively in connection with F and H tones. This is in line with the longer duration we have observed for the {F,H} toned vowels. A problematic finding for this link between diachronic lengthening and synchronic phonetic variation due to tone is the longer duration for R toned vowels. Diachronically, some words with R tones are shortened, not lengthened, at least in BKKT (but see [2] for lengthening effects of R tones in other Thai varieties). This issue brings us to the interactions between tones and the presence of codas.

In exp. 2, we studied short vowels combined with (sonorant) codas. We found the following differences among durational effects of tone on vowels: {R} < {M,L} < {F,H}. In the presence of codas, we observed that the R tone results in shorter vowel duration, compared to M and L, while H and F still results in the longest vowel duration, just like for long vowels in open syllables. Against the background of these findings, we note again that the order {M,L} < {F,H} is the reverse of the one reported in [3] and of what is predicted by an inversely proportional relation between average f0 height and vowel duration [2]. Moreover, the durational effects {M,L} < {F,H} can again straightforwardly be explained in terms of faster f0 falls and slower f0 rises. In particular, tones that comprise only an f0 fall (M, L) are shorter than other tones. The shortest duration of the R-toned vowels is, however, slightly surprising since this tone involves both an f0 fall and an f0 rise. This last finding suggests caution in attributing all effects purely to physiological considerations, like f0 control mechanisms. Phonological structure also seems to play a role. Whatever the exact explanation behind our findings, all of these effects closely mirror diachronic changes in BKKT vowel length [8]. The F and H tones, associated with the longest vowels, are known to correlate with diachronic lengthening; while the R tone, associated with the shortest vowels when codas are present, correlates with diachronic shortening. This is exactly what we observe diachronically with short and long vowels, respectively. Finally, we also observed interactions between tonal effects and coda types that also closely mirror diachronic changes. Lengthening of words with /-j/ and F tone is consistent with an overrepresentation of this combination among words that underwent diachronic lengthening. Similarly, shortening of words with /-w/ and R is a well-attested phenomenon diachronically in BKKT, e.g., [t^hɛ̌:w] > [t^hɛ̌w] ‘line, queue’.

Some limitations of this work and future avenues for research should also be discussed. First, we only examined the effects of tone on one vowel quality, /a/ both long and short. More vowel qualities should be studied to investigate the possibility that some effects may be vowel specific. For instance, mid vowels have been reported to be more prone to shortening effects in BKKT, accordingly, tonal effects on them may be different. Second, we reported on data from /Ca:/ and /CaG/N/ environments, a logical next step is to examine /Ca:G/N/ to further verify that short and long vowels behave the same in closed syllables. The extent to which other Thai varieties may or may not conform to the picture presented in this paper is also another avenue for future research. Finally, an investigation that focuses on individual differences in the strength of tonal durational effects may help shed further light on the rationale behind this phenomenon and its (non-)physiological underpinnings. Notwithstanding these limitations, the data presented in this paper suggest a consistent ordering for BKKT tonal durational effects {M,L} < {H,F}; while the effects of the R tone are dependent on syllable structure. Syllable structure also mediates the effects of other tones, like L and F, on vowel duration. The synchronic phonetic effects we observed are mostly in line with phonological vowel lengthening observed diachronically in BKKT. Given that the amount of lengthening due to tone is of moderate size, up to ~20%, we find it unlikely that phonological changes are purely the result of synchronic phonetic biases. However, the synchronic variation we have reported can reasonably be considered a precursor for sound change. The systematic phonetic biases present in the production of the vowel /a/ may lead it to move more and more towards /a:/ realizations, in association with certain tones and coda types, through iterations in the perception production loop. Eventually, this drift in duration space may lead to a new phonological length categorization for exemplars of this vowel diachronically.

5. Conclusion

In this paper, we have shown that contemporary BKKT vowel duration, for both long and short vowels, is systematically influenced by tonal categories. For long vowels in open syllables, we observed an order {M,L} < {F,H,R}. For short vowels in syllables with sonorant codas, we observed an order {R} < {M,L} < {H} < {F}, as well as interactions between tones and different coda types. This ordering of tonal effects on duration are both in contrast with previous reports [3] and problematic for the idea that vowel duration is inversely related to average f0 height [2]. We have argued that different tonal effects may be better understood if vowel duration is related to differences in the speed of f0 falls, which are faster, *vs.* rises, which are slower. In other words, longer vowel durations may be associated with tones that have a slower f0 rising component and shorter durations with tones that consist entirely of a faster f0 falling component. Crucially, these effects are not purely physiological, as they are also mediated by syllable structure and coda types. Finally, the durational effects we reported are in line with observed diachronic changes in phonological vowel length that are known to have taken place in the history of BKKT. Phonetic synchronic variation due to tonal effects on vowel duration may, thus, be considered fertile soil for phonological sound change diachronically.

6. Acknowledgements

We want to thank the members of Quantitative Methods for the Study of Language course taught at Chulalongkorn University in the Spring Semester of 2022 for their comments on the second experiment presented in this paper. We also want to thank Sireemas Maspong and Pittayawat Pittayaporn for their comments on an earlier draft of this paper. We also wish to thank three anonymous SST reviewers whose comments helped improving the quality of our work. Finally, it is our pleasure to acknowledge financial support from the Cognitive Science Program at Cornell University (FB), and from the Southeast Asian Linguistics Research Unit and the Faculty of Arts at Chulalongkorn University (TS).

7. References

[1] A. C. L. Yu, "Tonal effects on perceived vowel duration," in *Laboratory Phonology 10*, C. Fougeron, B. Kühnert, M. D'Imperio, and N. Vallée, Eds. De Gruyter Mouton, 2010, pp. 151–168. doi: 10.1515/9783110224917.2.151.

[2] J. Gandour, "On the Interaction between Tone and Vowel Length: Evidence from Thai Dialects," *Phonetica*, vol. 34, no. 1, pp. 54–65, 1977, doi: 10.1159/000259869.

[3] A. S. Abramson, *The Vowels and Tones of Standard Thai: Acoustical Measurements and Experiments*, vol. 28.2. Bloomington: Indiana University, 1962.

[4] E. Zee, "Duration and intensity as correlates of F0," *Journal of Phonetics*, vol. 6, no. 3, pp. 213–220, Jul. 1978, doi: 10.1016/S0095-4470(19)31153-2.

[5] Q.-M. Kong, "Influence of Tones upon Vowel Duration in Cantonese," *Lang Speech*, vol. 30, no. 4, pp. 387–399, Oct. 1987, doi: 10.1177/002383098703000407.

[6] J. J. Ohala and W. G. Ewan, "Speed of Pitch Change," *The Journal of the Acoustical Society of America*, vol. 53, no. 1, pp. 345–345, Jan. 1973, doi: 10.1121/1.1982441.

[7] J. Sundberg, "Maximum speed of pitch changes in singers and untrained subjects," *Journal of Phonetics*, vol. 7, no. 2, pp. 71–79, Apr. 1979, doi: 10.1016/S0095-4470(19)31040-X.

[8] P. Pittayaporn, "Sound changes in Thai vowel length: the significance with respect to modern spelling," Chulalongkorn University, 2016.

[9] D. Povey *et al.*, "The Kaldi speech recognition toolkit," *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*, Jan. 2011.

[10] Paul Boersma and D. Weenink, *Praat: doing phonetics by computer*. 2022. [Online]. Available: https://www.praat.org.

[11] D. M. Erickson, "A physiological analysis of the tones of Thai.," Ph.D. Dissertation, University of Connecticut, 1976.

[12] D. Erickson, "Thai Tones Revisited," *Journal of the Phonetic Society of Japan*, vol. 15, no. 2, pp. 74–82, 2011.

Declination-adjusted Normalisation of Cantonese Citation Tones

Phil Rose

Independent Researcher, Australian National University Emeritus Faculty

https://philjohnrose.net

Abstract

An experiment is described to see if the normalisation of citation tone fundamental frequency (F0) can be improved by taking into account its occasion-specific decay. An optimum baseline z-score normalization of the F0 of ten Cantonese speakers' unstopped citation tones without such adjustment gave about a twenty-fold reduction in the between-speaker tonal variance (normalisation index = 21.3). It is shown that adjusting for F0 decay by simply using the linear slope of the phonologically level tones can improve on this baseline normalization a little, by up to about 11% (NI = 23.6). The result is also a representation closer to the tonal pitch.

Index Terms: normalisation, tonal F0, tonal pitch, F0 declination, Cantonese

1. Introduction

1.1. Declination

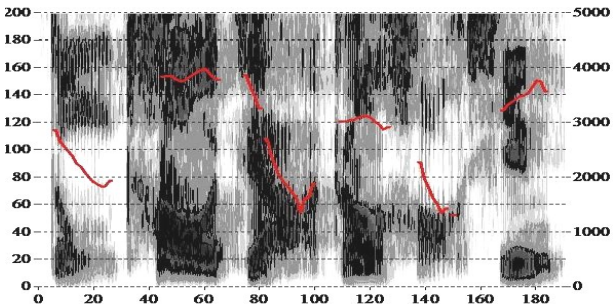


Figure 1: *Alternating sequence of L and H tones in a six-syllable Mandarin utterance (F0 superimposed on wideband spectrogram). Y axis-left = F0 (Hz),Y axis-right = spectral frequency (Hz), X-axis = duration (csec.)*

Fundamental frequency (F0) tends to gradually drop through an utterance: this is a phonetic commonplace [1 p. 53]. This so-called *declination* can be most easily seen and modeled in tone languages, where the language's phonology makes clear which of the utterance's F0 values are tonologically equivalent [1 pp. 69-71, 2 pp. 296-297]. This is illustrated in figure 1 with a Mandarin utterance 董超假装好心 *Dǒng Chāo jiǎzhuāng hǎoxīn* 'Dong Chao pretended to have a kind heart'. The utterance was taken from a recorded narrative [3 p. 116] which happened to have six syllables carrying a sequence of alternating L and H tones (tones 3 and 1). Figure 1 plots the F0 time-course for the utterance superimposed on a wide-band spectrogram to show the relationship of the F0 to the segmental structure. It can be seen that the falling F0 trajectory of the L tones on the odd syllables (which may be an extrinsic part of the tone's underlying fall-rise target or an intrinsic effect of the voiceless obstruent Onset consonants) gradually decreases through the utterance. The higher F0 of the H tones also decreases over the first two even syllables. The higher value of the F0 on the H tone of the final syllable *xīn* [ɕɪn] probably reflects intrinsic vowel effects, as the utterance was played to a native speaker phonetician who said that it was unmarked and the final syllable not perceptually prominent.

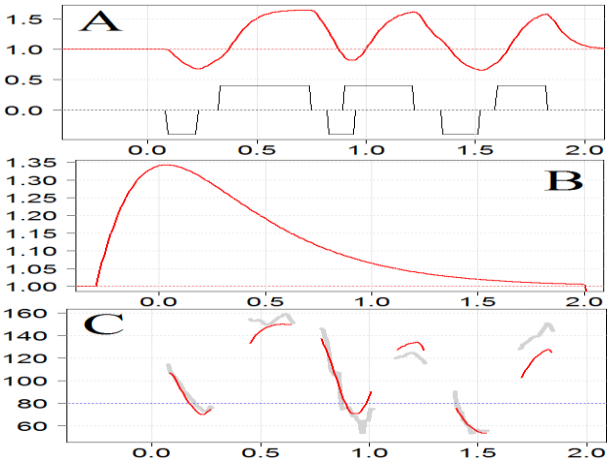


Figure 2: *Declination decomposition with Fujisaki command-response model of F0 in the utterance in figure 1. A = Tonal impulses and response, B = Phrasal Response, C = actual (grey) and estimated (red) F0. X-axis = duration (sec.).*

Declination has been widely studied from several aspects, including to what extent it is under a speaker's control and is used to convey linguistic and paralinguistic information; and how best to model it. A useful overview is in [1]. The well-known Fujisaki *command-response* model, e.g. [4 5] provides an insightful way of separating out the declination from the tonal components. This model factors the time-varying F0 into two types of component, both modeled as impulses of given amplitude and duration. A *tonal* component represents the response of the speech production mechanism to impulse commands for implementing tone. Such impulses are shown on the bottom line of panel A of figure 2. They consist of a string of equal amplitude impulses of alternating polarity. These model the low and high tones respectively [6] and are said to relate physiologically to the *pars recta* activity of the crico-thyroid [7, p.4]. The impulse response to these commands of differing polarity is shown in the top line of panel A in red. The second, or *phrasal*, component represents a much slower time-varying response and accounts for the gradual change in F0 – the declination – throughout an utterance or international phrase. Its response is shown in panel B. This is said to correspond physiologically to the *pars obliqua* of the crico-thyroid, but might also represent gradually decaying sub-glottal pressure. Panel C shows the combined tonal and phrasal responses in red and the actual F0

in grey. It can be seen there is a fairly good fit. A better fit could of course have been obtained if the amplitudes of the individual tone impulses had been individually manipulated, but that would have obscured the point of factoring the F0 into a fixed tonal and declinational component (the duration of the tonal components, however, still had to be independently specified to capture the metrical structure of the utterance).

1.2. Citation tone declination

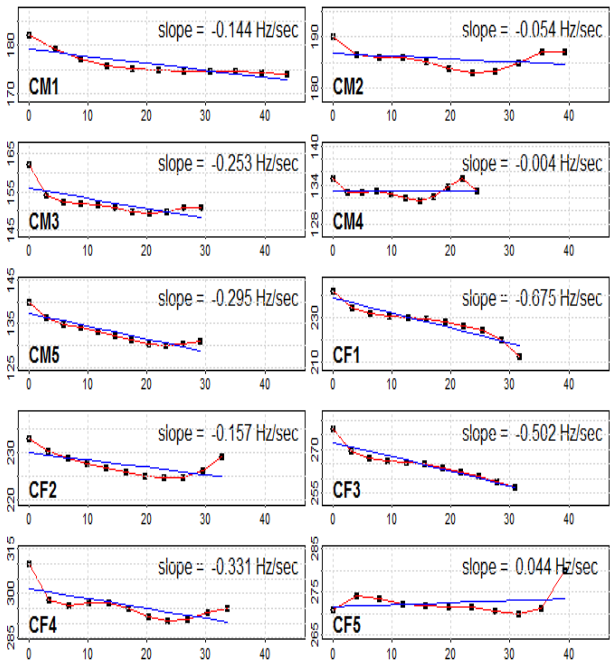


Figure 3: *Mean F0 as function of mean duration for Cantonese high level tone in five male (CM 1 – 5) and five female (CF 1-5) speakers. Blue line = least squares regression. X-axes: mean raw duration (csec.), y-axis = mean F0 (Hz.).*

A gradual decrease in F0 can also be sometimes observed on tones in monosyllabic utterances, as for example when citation tones are elicited. Figure 3 shows F0 plotted as a function of duration for five female and five male Cantonese speakers' high level citation tone. A least-squares regression line has been fitted, and the value of its slope shown. (The data are from a multispeaker acoustic description of conservative Cantonese citation tones [8], where the tones were read out in random order from Chinese characters on prompt cards. Each speaker's F0 shape is the mean of 16 tokens balanced for intrinsic vowel F0 and sampled at 10% points of duration. The mean F0 and duration data may be downloaded from [9].

It can be seen firstly that speakers differ in their decay, but that, overall, males have less F0 decay than females. Since the F0 of males is usually lower than females, this suggests that the rate of decay may be related to overall F0. But there are clearly within-speaker differences that do not relate to overall F0. CM4 and CM5, for example, differ the most in F0 decay but have very similar F0 for their high level tone; and CF1 and CF3 have similar decay but rather different F0. Importantly, there is also no obvious correlation of F0 decay with duration (one might expect that the longer the tone were sustained the further its F0 would drop). This suggests that the F0 decay may be occasion-dependent, and that adjusting for this may help to improve tonal F0 normalization. It is the purpose of this paper to investigate this hypothesis using citation tonal data from Cantonese.

2. Normalisation

An even more basic phonetic commonplace than declination is that speech acoustics inevitably bear the imprint of the individual vocal tract that produced them, as well, of course, as the brain that drove that vocal tract. If we are focusing on the speech of the individual, as for example in forensic voice comparison [10], then this is indeed desirable. But if our focus is *language*, then it is often necessary to remove as much speaker-dependent acoustic material as possible so as to arrive at a quantified parametric representation of the variety under question. This in turn is necessary for many important dialectological, socio-phonetic, typological and even historical applications, such as quantifying the tones of a variety [11], comparing varieties with respect to their tones [12] or even reconstructing tonal acoustics [13].

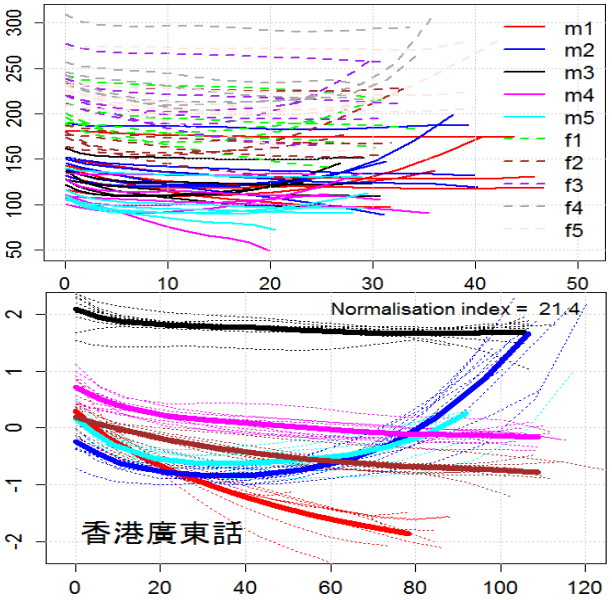


Figure 4: *Normalisation of Cantonese unstopped citation tonal F0. Top = raw mean tonal F0 for 10 speakers' tones. Dashed lines = females. X-axis = mean duration (csec.), y-axis = mean F0 (Hz). Bottom panel = z-score normalised F0. Thicker lines = mean normalised F0. X-axis = normalised duration (%), y-axis = normalised F0 (sds around mean).*

Figure 3 shows the normalisation of the tones of the same 10 Cantonese speakers shown in figure 3. On syllables ending in a sonorant, conservative Hong Kong Cantonese contrasts six tones: three with level pitch, two with rising pitch, and one with falling pitch. (The term *pitch* is used here in its proper perceptual sense, not as a synonym of F0.) The three level-pitched tones are located at the top, in the middle and just below the middle of the speaker's pitch range. Examples are [si: high level pitch] *poem* 詩; [si: mid level pitch] *try* 試, [si: lower-mid level pitch] *event* 事. Both rising tones start low in the pitch range, with one rising to high and one to mid: [si; low pitch rising to high] *shit* 屎, [si: low pitch rising to mid]

市 *market*. The falling tone starts low and falls still lower, such that its phonation type usually becomes non modal (breathy or creaky) as it falls below the speaker's normal pitch range: [si: low falling pitch] *time* 時.

The top panel of figure 4 plots the 10 Cantonese speakers' six tones' raw mean F0 trajectories as a function of raw mean duration. Apart from the unsurprising fact that the females' tonal F0 generally lies, with an overlap, higher than the males', the result is rather a mess: it is difficult to see from this figure how many tones there are and what their F0 trajectories are like. The bottom panel of figure 4 shows a z-score normalisation of the data [14] which resolves the raw tonal F0 nicely into six groups. It was shown in [15] that this type of normalisation easily outperforms other types of normalisation proposed. The performance if the normalisation in figure 4 was therefore used as a baseline.

So the question this paper asks is: can an improvement on this baseline normalization be made by incorporating the type of F0 decay evident in figure 3? Although the z score normalisation is well-known, its evaluation may not be. The following section thus addresses the numerical evaluation necessary to determine an improvement in normalisation.

2.1. Numerical evaluation of normalisation

The effectiveness of normalisation is currently estimated by the method used in the first tonal normalisation study of some decades ago, on Vietnamese tones [16, p.133ff.]. Variance plays a crucial role. Before normalisation the between-speaker variance in raw tonal F0 values will tend to be large because of between-speaker differences in tonal F0 caused by between-speaker differences in mass and length of the vocal folds. A female's high tone may have twice the F0 of a male, for example: this effect of secondary sexual bimorphism can be seen in figure 4. After normalisation it is hoped that the between-speaker differences in tonal values will be minimised. Consequently, evaluation of the normalisation strategy involves quantifying how much the normalisation reduces the between-speaker tonal variance in the unnormalised data, a quantity called the *normalisation index* (NI). The idea is to estimate, for both raw and normalised data, the proportion of the overall variance in the data that is due to the between-speaker variance within tones. This is called the *dispersion coefficient*. Since the point of normalisation is to minimise between-speaker differences in tones, the proportion of the overall variance that is due to between-speaker tonal differences is expected to be smaller after normalisation, and so the ratio of the dispersion coefficients for the raw and normalised data – the *normalisation index* – quantifies by how much the between-speaker tonal variance has been reduced and how much between-speaker differences in tonal F0 have been minimised.

Using the 10 Cantonese speakers' data in the top panel of figure 4 as an example, the calculation of the NI can be formulated thus. Let $F0_{ijk}$ be the F0 value for the i^{th} speaker's j^{th} tone at the k^{th} sampling point. In the Cantonese data for example, i = 1… 10 speakers; j = 1… 6 tones; and k = 1… 12 sampling points (0%, 5%, 10% 20% … 100%). Then the mean F0 value over all speakers at a given sampling point in a given tone $\overline{F0}_{.jk}$ is:

$$\overline{F0}_{.jk} = \frac{1}{10}\sum_{i=1}^{10} F0_{ijk} \qquad (1)$$

The variance around the mean F0 value over all speakers at a given sampling point in a given tone $S^2_{\overline{F0}.jk}$ is:

$$S^2_{\overline{F0}.jk} = \frac{1}{10}\sum_{i=1}^{10}\left(F0_{ijk} - \overline{F0}_{.jk}\right)^2 \qquad (2)$$

The mean of the variances $S^2_{\overline{F0}.jk}$ at all 12 sampling points of all tones, called *between-speaker tonal variance* $\overline{S}^2_{\overline{F0}.jk}$ is taken as an estimate of the variance representing between-speaker differences in tonal values:

$$\overline{S}^2_{\overline{F0}.jk} = \frac{1}{72}\sum_{j=1}^{6}\sum_{k=1}^{12} S^2_{\overline{F0}.jk} \qquad (3)$$

For the raw Cantonese data in the top panel of figure 4 this was 2580.0. In order to quantify the proportion of the overall variance taken up by variance associated with between-speaker differences in tone, the between-speaker tonal variance is then normalised with respect to the overall variance of the data. This is the mean of the between-speaker variances at each sampling point, i.e. ignoring the tonal differences. For the raw Cantonese data, this was ca. 2887.9. The ratio of the between-speaker tonal variance to the overall variance is called the *dispersion coefficient* (DC). In this case its value of (2580.0 / 2887.9 =) ca. 89.3% indicates that there is almost as much variation *between* the Cantonese speakers' raw tonal values as in the data overall, and that they effectively do not cluster.

Since normalisation is intended to reduce the between-speaker differences in tonal F0, one expects the DC for the normalised data to be substantially smaller than the DC for the raw data. It is calculated, *mutatis mutandis*, in the same way as the raw DC, namely as the ratio of *between-speaker normalised tonal variance* to *overall normalised variance*. The DC for the normalised Cantonese data was (0.04 / 0.94 =) ca. 4.2%, indicating that only a small amount of the overall variance was taken up by between-speaker differences in tone. The normalisation index (NI) is then defined as the ratio of normalised DC to raw DC. For this normalisation, the NI was (89.3% / 4.2%) = ca. 21.3, meaning that normalisation has resulted in about a twenty-fold reduction in the proportion of variance in the raw data due to between-speaker differences in tone.

3. Declination-Adjusted Normalisation

3.1. Procedure

For non-rising tones, negative and positive F0 offset perturbations can obviously effect estimation of the slope. In order to control for this, the F0 at the last two sampling points in these tones was removed. The remaining raw F0 trajectory was then modeled with a 4th degree polynominal to permit resampling of F0 at 0%, 5%, 10%, 20%, …, 100% of duration. A least-squares regression line was then fitted to the F0 trajectory, and the raw F0 adjusted by its slope and intercept according to (4)

$$F0_t' = F0_t - (mt + b) \qquad (4)$$

where $F0_t'$ = declination-adjusted F0 at time t, $F0_t$ = mean F0 at time t, m, b = slope, intercept of least squares regression line, t = time t. This process is illustrated in the left panel of figure 5, which shows how the tonally relevant F0 for CF1's high level tone (solid grey line) is adjusted by the slope of its least squares regression line (black dotted line) to its declination-adjusted value (solid red line).

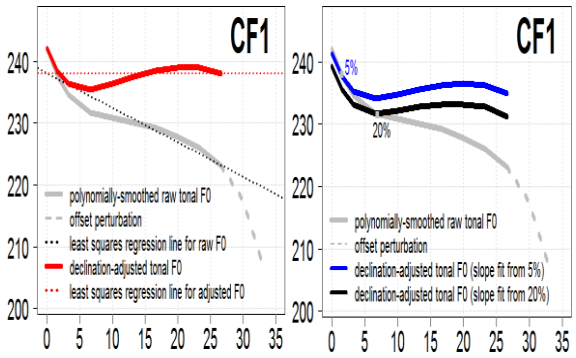


Figure 5: *Illustration of F0 declination adjustment for Cantonese female 1's high level tone. = X-axis = duration (csec.) Y axis = F0 (Hz).*

3.2. Parameters and results

Two parameters were manipulated in the declination-adjustment. The initial few centiseconds of all speakers' F0 show consonantally-induced onset perturbations, and, as these will affect the tone's F0 slope, normalisation was tested with increasing portions of initial F0 trajectory (0%, 5%, 10%, 20%) removed to estimate the slope. Allowing the onset to vary from 0% to 20% of duration tests whether a better normalization is achieved if the onset perturbation is not included in the F0 slope. This is illustrated in the right panel of figure 5, which shows the adjusted F0 for CF1's high level tone when the slope is based on the raw F0 with the first 5% removed (blue line), and with the first 20% removed (black line). Because of the shape of the initial part of the raw tonal trajectory, removing these portions gradually decreases its slope. As can be seen in figure 5, the effect of this is to pivot the declination-adjusted F0 around the percent sampling point onset. This parameter is therefore called *pivot*, and slopes were generated using four pivot values at 0% 5% 10% and 20% of the raw tonal duration.

Table 1: *Normalisation indices from different normalisation trials.*

pivot	0%	5%	10%	20%
option 1	slope from high-level tone			
	23.0	**23.1**	23.0	22.4
option 2	slope from mid-level tone			
	23.1	22.8	22.4	22.1
option 3	slope from lower-mid-level tone			
	21.6	21.5	21.2	20.8
hybrid combinations				
option 4	23.5	**23.6**	23.3	22.9

It would, with six different tones, be possible to adjust each tone by a different slope. Apart from the overfit that this is likely to cause (if you torture the data long enough, it will confess!), it is also counterintuitive to imagine six different declination factors at play; and in addition a declination slope can only reasonably be estimated from a citation tone which can be assumed to have a level pitch target (of which there are three in Cantonese). Therefore declination-adjusted normalisation was first tested with three options, where all tones were adjusted by a single slope estimated from each of the three level tones (high-level, mid-level, lower-mid-level). Results are shown in table 1, where it can be seen, firstly, that the pivot does have a slight effect. Generally the performance decreases with increasing amounts removed from the initial part of the tone trajectory: apparently it is not a good idea to remove onset perturbations. As far as between-tone differences are concerned, the slope from the lower-mid-level tone (option 3) performs the worst, and only marginally better than the baseline value of 21.3. The high-level tone slope (option 1) outperforms the mid-level tone slope (option 2) for three out of four pivots, although both achieve the same maximum normalisation index of 23.1. This is a small improvement on the baseline value of 21.3, indicating that adjusting for declination can marginally improve clustering.

It was also tested whether a better performance can be achieved with a hybrid approach using both slopes from the high- and mid-level tones to separately adjust different sets of tones (option 4). It was found that when all tones except the mid-level tone were adjusted by the slope of the high-level tone, and the mid-level tone was adjusted by its own slope, a slightly better NI of 23.6 is obtained – an improvement over the baseline of about 11%. This suggests that F0 decay and F0 height may be related in a more complex way such that slope may increase with distance from a central F0 value. This is an obvious thing to test.

Figure 6 shows the resulting normalisation with option 4. It can be seen that the level tones now have correspondingly level normalised F0 over much of their duration.

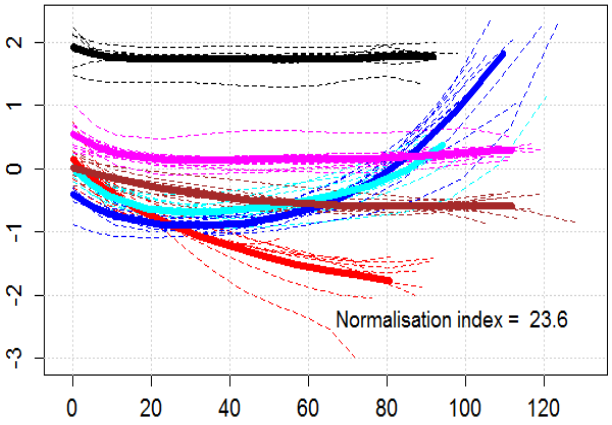


Figure 6: *Declination adjusted z-score normalisation for Cantonese unstopped tones = X-axis = normalised duration (csec.) Y axis = normalised F0 (sds).*

4. Summary

This paper has used two phonetic commonplaces – F0 declination and between-speaker differences in acoustic output – to show that normalisation of tonal F0 can be improved a little by taking into account, with very simple modeling, what is probably occasion-specific decay in tonal F0. Whether this is appropriately called declination is moot. The next questions to address are whether further improvement can be achieved by more accurate modeling of the decay – perhaps with Fujisaki parameters – and how to estimate an appropriate slope in tonal systems (like Shanghai) that lack tonologically level tones.

5. Acknowledgements

Very many thanks to my three anonymous referees whose comments let me see where I was not making myself adequately clear. I have tried where possible to incorporate their advice.

6. References

[1] D.R. Ladd, "Declination: a review and some hypotheses," *Phonology Yearbook*, vol. 1, pp. 53–74, 1984.

[2] P. Rose, "Where the science ends and the law begins: likelihood ratio-based forensic voice comparison in a $150 million telephone fraud," *Int'l Journal of Speech Language and the Law*, pp. 277–324, 2013

[3] Editing Team, 水滸傳 *Water Margin*, Mandarin Audio-Visual and Resources Centre of Hong Kong, no date.

[4] H. Fujisaki, "Dynamic Aspects of Voice Fundamental frequency in Speech and Singing," in F. MacNeilage [Ed], *The Production of Speech*, pp. 39–55, Springer, 1983.

[5] H. Fujisaki, S. Ohno, and W.T. Gu, "Physiological and Physical Mechanisms for Fundamental Frequency Control in Some Tone Languages and a Command Response Model for generation of Their F0 Contours," International Symposium on Tonal Aspects of Languages, Beijing, 2004.

[6] H. Fujisaki, K. Hirose, P. Hallé, and H. Lei, "Analysis and modeling of tonal features in polysyllabic words and sentences of the Standard Chinese," in *ICSLP 1990 - 1st Int'l Conf. on Spoken Language Processing, Kobe, Proceedings, 1990*, pp. 841–844, 1990.

[7] H. Fujisaki, "In Search of Models in Speech Communication Research," in *INTERSPEECH 2008 – 9th Annual Conference of the International Speech Communication Association, Brisbane, Proceedings*, 2008, pp. 1–10.

[8] P. Rose, "Hong Kong Cantonese Citation Tone Acoustics: A Linguistic-Tonetic Study', in M. Barlow [Ed] *SST 2000 – 8th Australian Int'l Speech Science and Technology Conference*, *Canberra, Proceedings*, 2000, pp. 198–203.

[9] https://philjohnrose.net/numerical_data/index.html

[10] P. Rose, "Likelihood ratio-based forensic voice comparison with higher level features: research and reality," in E. Lleida & L. J. Rodriguez-Fuentes [Eds] *Recent Advances in Speaker and Language Recognition and Characterisation*, pp. 476–502, *Computer Speech and Language,* Special Issue, 2017.

[11] P. Rose, "A Linguistic-Phonetic Acoustic Analysis of Shanghai Tones", *Australian Journal of Linguistics*, vol. 13, pp. 185–220, 1993.

[12] W. Steed and P. Rose, "Same tone, different category: linguistic-tonetic variation in the areal tonal acoustics of Chu-qu Wu", in *INTERSPEECH 2009 - 8th Annual Conference of the International Speech Communication Association, Brighton, Proceedings*, 2009, pp. 2295–2298.

[13] P. Rose, "Oujiang Wu tones and Acoustic Reconstruction," in C. Bowern, B. Evans, and L. Miceli [Eds.], *Morphology and Language History,* pp. 235–250, John Benjamins, 2008.

[14] P. Rose, "Some considerations in the normalisation of the fundamental frequency of linguistic tone," *Speech Communication*, vol. 6, no. 4, pp. 343–352, 1987.

[15] P. Rose, "Comparing Normalisation Strategies for Citation Tone F0 in Four Chinese Dialects," in C. Carignan & M. D. Tyler [Eds], *SST 2016 - 16th Australasian Int'l Conf. on Speech Science & Technology*, *Sydney*, 2016, pp. 221–224.

[16] M.A. Earle, *An acoustic phonetic study of North Vietnamese tones*, Monograph 11, Speech Communication Research Laboratories Inc., Santa Barbara, 1975.

A Machine Learning Ensemble to Automatically Classify Tongue Ultrasound Contours Based on Displacement Measurements

Simon Gonzalez

The Australian National University

u1037706@anu.edu.au

Abstract

This paper introduces a Machine Learning Ensemble Model to automatically classify tongue ultrasound contours. It has been trained on displacement measurements in English Coronal Obstruents (/t/, /s/, /tʃ/, /ʃ/), from eight female native speakers of Australian English. The model has an accuracy of 97.6% for the Random Forests and 74.4% for the Decision Tree. The accuracy is higher for fricatives than for stops. Results also show that the most reliable area for classification is from 20% to the 40% of the contour length, which corresponds to the tongue area between the tongue front and the tongue body.

Index Terms: speech ultrasound, tongue contours, Machine Learning, English coronal obstruents, displacement

1. Introduction

The tongue is described as a highly mobile organ, whose shape can be deformed in extremely rapid succession through subtly different movements [1]. One of the main reasons for these changes in shape is to achieve articulatory targets. This complex motions and intricate postural adjustments make it difficult to classify tongue shapes during speech [2]. One of the technologies that has been used for this purpose is tongue ultrasound imaging (henceforth ToUS), which is used to identify contours from images as well as articulatory landmarks [3][4][5][6].

Due to the time-consuming aspect of tracing and extracting contours from ToUS images, several automatic and semi-automatic approaches have been developed for identification [7][8] and error detection [9]. Since the rise of Machine Learning (henceforth ML) in the last two decades, it has been one of the most widely implemented techniques to achieve automatic tasks in ToUS imaging. These techniques include Deep Learning [10], Support Vector Machines [2], Deep Neural Networks [11], Deep Belief Networks [12], and Convolutional Networks [13][14].

These implementations have made invaluable contributions to the tasks of contour identification and extraction, and they have increased the amount of data that can be processed in unprecedented ways. However, having larger datasets demands adequate computational approaches that can be developed for efficient processing and analysis of articulatory data [15]. This analysis area is one where ML has not been implemented, especially in ToUS imaging for linguistic purposes. One relevant question is whether ML models can offer more efficient ways to deal with the complexity of ultrasound data and its articulatory implications. One of the challenges in ToUS analysis is that identifying relevant articulatory sections of the tongue is not always a straight-forward process. For example, when investigating the articulation of alveolar segments in English, the front section of the tongue is a relevant area to observe, but there are other sections of the tongue, e.g. the tongue body, which also shows strong articulatory activity [16]. In this case, the question is not just what happens at the front section, but also at other areas that can be relevant for the gestural description of the segments, information that can be used to automatically classify contours to the right category.

Current analyses of ToUS include SSANOVAs [17][18][19], distances [20][21], velocities [22][23][24], heatmaps [16][25], and Principal Component Analysis [26][27]. These methodologies have been able to address relevant linguistic questions, providing strong analysis frameworks. However, much of the process of analysis still requires a great level of qualitative interpretation on behalf of the researcher. This is an accurate approach, yet not very efficient when large datasets are analysed. Another issue with these approaches is that they generally analyse data based on individual speaker patterns, then generalisations are made across all speakers in the data. What is therefore needed is approaches that can analyse all available data and give all factors (speaker-dependent and speaker-independent factors) the same opportunity to contribute to the classification and description of tongue contours. This is the reason why the implementation of ML can offer an opportunity to examine ToUS data in a more generalisable way.

2. Aim of Paper

The aim of this paper is to develop and implement an ML ensemble approach to analyse and automatically classify mid-sagittal tongue contours of English coronal obstruents, specifically palato-alveolars and alveolars. This is designed to capture spatio-temporal tongue gestures, the dynamics of the tongue contour over time, from previous vowel to maximum constriction. We employ a grid-lines approach, where tongue displacements between articulatory landmarks are captured by multiple lines and create descriptors to classify tongue trajectories across articulatory landmarks. We aim to train an ML ensemble (Random Forests and Decision Trees) using paired data to predict new data. We chose displacement measurements as the baseline for classification. These are calculated as tongue movements between two landmarks, previous vowel and maximum constriction for each of the four segments analysed, two alveolars and two palato-alveolars.

3. Methodology

In this section, we describe the methodology used for the analysis. We first present information on the data (equipment, collection, processing), then on the acoustic and articulatory landmarks, and finally on the measurements and the ML ensemble developed.

3.1. Equipment and Data Formats

We recorded mid-sagittal images of the tongue contours using a portable Sonosite 180 Plus ultrasound machine with a C11/7-4 MHz 11-mm broadband curved array transducer. The transducer was fixed to a stabilisation helmet from Articulate Instruments [24]. The acoustic signal was recorded using a Shure KSM137. The microphone was connected to an M-audio DMP3 preamplifier, and the audio output from the amplifier was sent to the audio input of a camcorder. A Sony DCRTRV103 digital camcorder in NTSC format (30 fps) did a simultaneous recording of both ultrasound and acoustic signals. The data was then downloaded to a computer using the Adobe Premiere Elements software (www.adobe.com). The audio signal was digitised at a sampling frequency of 48 KHz with 16-bit quantisation. Each video was then saved as a sequence of still images in JPG format (29.97 fps). We also saved the audio signal as WAV files. The audio recordings were saved at a sampling frequency of 44.1 KHz with 16-bit quantisation.

3.2. Speakers and Stimuli

The participants were eight adult females, all native speakers of Australian English with no reported hearing or articulatory impairment. Target segments were two alveolars (/t/, /s/) and two palato-alveolars (/tʃ/, /ʃ/). The target segments were elicited in monosyllabic words in onset position: *tack*, *sack*, *Chack*, and *shack*. The vowel /æ/ (the lowest front vowel in Australian English, see [28]) was chosen as the common context vowel because of its articulatory properties, being the most suitable context vowel for an ultrasound study of coronal segments. The target words were placed in a carrier sentence. This isolates relevant articulatory parameters for the segments analysed, in terms of tongue advancement and tongue height. The carrier sentence was *Please, utter X publically*, where *X* represents the target word. This controls the previous vowel /ə/. In Australian English, a non-rhotic variety, the word utter is pronounced [ˈʁrə]. Thus, the target segment occurred after the mid central vowel /ə/ and before the vowel /æ/.

3.3. Data Segmentation

The still images in JPG format were used for tracking the tongue surface lines (contours) using the EdgeTrak software [29]. The audio signal was used for the acoustic analysis of the data. Contours were saved as con files, which is the default format in Edgetrak, and they are coded as numeric Cartesian values for each point of the contour line ([*x,y*] coordinates), and totalling 100 points per contour. For each of the four segments, we selected five repetitions. The format of the data follows the structure as in [16] with a csv file containing the following columns: Speaker, Segment, Repetition, and Frame.

3.4. Landmark Identification

We identified two landmarks: previous vowel (PV) and maximum constriction (MC). The motivation is to examine which parts of the tongue are the ones that activate and make the main articulatory gestures to achieve the constriction. The identification of these landmarks was based on two approaches. PVs were labelled by the authors from spectrogram and waveform display in Praat [30] (further description in Section *3.4.1*) and MCs based on articulatory landmarks using the Ultrasound and Visualisation App (UVA) [16], which carries out both static and dynamic analyses from the numeric output files exported by EdgeTrack. The app has a landmark identification task, in which users visualise contours in context (frames before and after), which helps the coding of landmarks. We used this functionality to assign the corresponding Maximum Constriction (MC) contours for each sequence. Below, we present the acoustic and gestural cues observed when identifying the landmarks.

3.4.1. Acoustic identification of Previous Vowel /ə/

The vowel /ə/ was segmented on the basis of clear vocalic peak pulses in the acoustic signal. The onset was placed at zero crossing point preceding the peak of the first clear vocalic pitch pulse (boundary (a) in Figure 1) after the offset of /t/. The offset of /ə/ was placed at zero crossing point following the peak of the last clear vocalic pitch pulse (boundary (b) in Figure 1) before the onset of the following target consonant. The tongue contour corresponding to the time mid-point of the vowel duration was assigned as the PV for each sequence.

3.4.2. Articulatory identification of Maximum Constriction

MC contours were identified by holistically examining ToUS contours one by one in each sequence. The MC frame was the frame previous to the first downward movement of the tongue from consonant MC to the following vowel /æ/. This was the contour in which the maximum raising/advancement was observed.

3.5. Measurements

The analysis baseline was done on a gridlines approach, which is implemented within UVA and also in other work (c.f. [23][31][32]). With this type of approach, we can carry out ToUS analysis in both advancement and height dimensions, across all the extent of the contours available. The process is represented in the Figure 1. The first step is to create a point of origin from the available contours. Then gridlines are projected from this point to capture sections relevant for articulation. Then intersections are calculated between the gridlines and the contours, which gives the common area of analysis encompassing all data that can be compared across all available contours. In our analysis, we selected 20 gridlines to capture the necessary articulatory activity for both upward and downward movement of the tongue. This is done for each speaker individually.

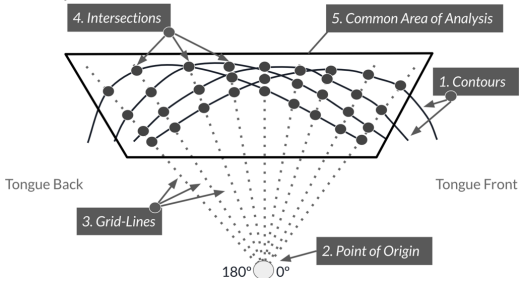


Figure 1: *Gridlines and intersections calculations.*

Since the gridlines are fixed for all contours in a given speaker, then any observed difference can be reliably interpreted as pertaining to articulatory differences.

We then selected the dynamic measurements and worked on the displacement calculations in mm as measurement units. Here, motion is defined as the change of a tongue section within the mouth in respect to time, more specifically, how fast the tongue is moving. Displacement, which is the length of the path travelled by the tongue section from one landmark to another, based on a specific gridline. It is important to point out that this

displacement is a relative measure from one tongue contour to another, i.e. the displacement calculated can only measure the relative movement from point A to point B in a given gridline.

The process is represented in Figure 2, and it first calculates the distances from the origin point to all contour intersections across all gridlines. For each gridline, new distances are calculated from the first landmark (PV) to the second one (MC). The displacement can therefore be positive (e.g. first gridline), zero (e.g. second gridline), or negative (e.g. third gridline).

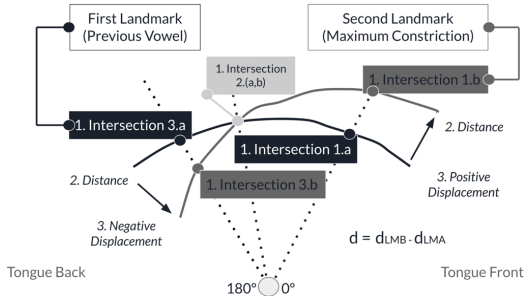


Figure 2: *Displacement calculation between landmarks.*

3.6. Analysis

For the analysis, we used a supervised ML algorithm ensemble in R [34]. An important characteristic of supervised algorithms is that they are trained to identify the category to which an object belongs based on the characteristic of the object itself [33]. We then divided the data into training (70%) and test (30%) subsets.

This approach was chosen to address two main goals. The first one was to identify which of the sections across the tongue contour surfaces were the ones relevant to describe the distinctive gestural behaviours. In order to keep measurement consistency, we grouped the 20 gridlines into five percentage incremental sections from right to left: gridlines 1 to 4 (0%-20%), 5-8 (20%-40%), 9-12 (40%-60%), 13-16 (60%-80%), and 17-20 (80%-100%). Lower gridlines (and percentages) would capture more advanced articulatory gestures at the front section of the tongue, and higher gridlines (and percentages) would capture more retracted gestures at the back sections of the tongue.

To address this research goal, to identify and distinguish relevant articulatory sections, we used a Random Forest algorithm (RF), which combines the output of multiple decision trees to reach a single result. RF is employed for classification, regression, and other activity based on the construction of a multitude of decision trees during the training and generates the class that represents the overall prediction of the single trees [35].

The second goal was to identify the amount of displacement that was relevant when classifying tongue contours. For this, we ran a Decision tree algorithm, which creates a model to predict the value of an outcome variable by learning decision rules from data features [36]. We used this to examine which tongue contour percentages are used to distinguish between the type of segments, or categories (alveolars vs palato-alveolars) and at what stage in the classification process.

4. Results

For both ML algorithms, we ran models where the category to predict was the segment individual type. We added the five percentage sections and the speaker as predictors. The reason to keep speaker as a predictor was to check whether individual speaker articulatory patterns or individual displacements were relevant. The raw displacement values for all speakers are shown in Figure 3 below. These show that most of the distinctive articulatory patterns happen in the first 60%-70% percent of contours. For the 80%, the differences are smaller than the other percentages, and in the 100% all displacements are negative. These indicate that all segments lower the back of the tongue at 100% and the 80% functions as a pivot or anchor in the tongue movement (see [37] for similar results in vowels).

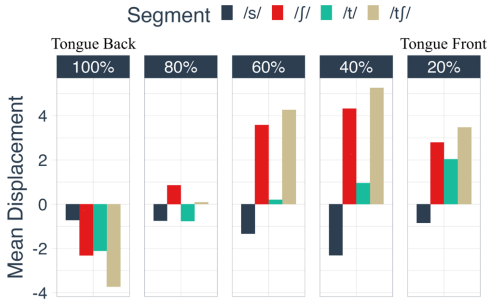


Figure 3: *Raw displacement across segments.*

4.1. Random Forest Results

The RF classification task had an accuracy of 97.6%. The only four wrong classifications came from four /ʃ/ tokens which were classified as /tʃ/. The variable importance results from the RF model are shown in Figure 4. These show that most of the distinctions across all segments take place at the 40%, which corresponds to the area of the tongue between the tongue front (highly active in alveolars) and the tongue body (highly active in palato-alveolars).

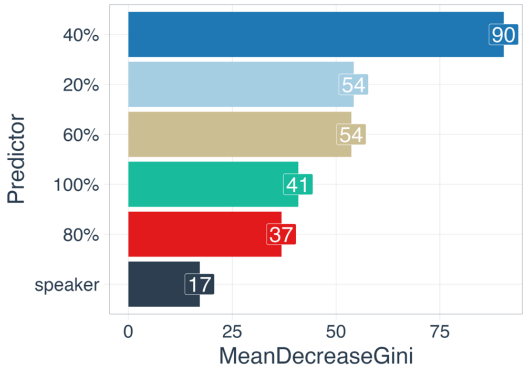


Figure 4: *Variable Importance from the Random Forest Model.*

4.2. Decision Tree Results

Figure 5 shows the Decision Tree classification of the test data from the training data. The figure shows the parameters that are most relevant for the classification of segments. It shows that the 40% is the most relevant parameter, since it is used several times and at top splits to distinguish between categories. This is similar to the RF results, but it adds further details by specifying at what point the distinctions are activated. Results also show the displacement thresholds at each of the nodes. It shows that at Node 1, there is a clear distinction between alveolars and palato-alveolars at the 40% and showing that displacements below 1.2mm are a cutoff point, with alveolars having lower displacements and palato-alveolar higher ones. Nodes 2 and 3 separate /s/ from /t/, with /t/ having higher displacements than /s/ at 40% and 80%. Nodes 4 and 5 separate /tʃ/ from /ʃ/, at both 80% and 100%. At the 80% section, /ʃ/ has lower displacements

that /tʃ/, with the cutoff being 3.1mm. Examining the final classifications of the tree, results show that when distinguishing between manner of articulation at the same place of articulation, stops display higher displacements than the fricative counterparts.

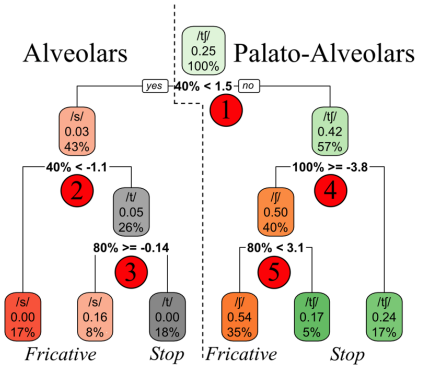


Figure 5: *Decision Tree from the trained model.*

The Decision Tree model had a predictive accuracy of 74.4%. The prediction accuracy is shown on Figure 6, which combines a correlation heatmap plot and a dendrogram. Segment /s/ was the segment with the highest accuracy (90%), with 6% classified as /ʃ/ and 4% as /t/. Segment /ʃ/ had the second highest accuracy (80%), with 12% classified as /tʃ/ and 7% as /s/. /t/ had the lowest accuracy (59%) with 33% classified as /ʃ/ and 9% as /s/. Finally, /tʃ/ had slightly higher accuracy (67%), with 33% classified as /ʃ/. These general results show that fricatives have higher accuracy than stops, 85% and 63%, respectively.

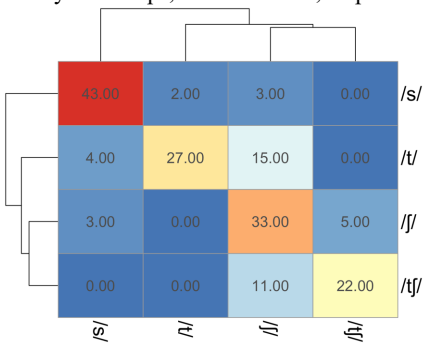


Figure 6: *Accuracy and errors in the Decision Tree output.*

5. Discussion

Random Forest results show the order of importance of the sections that are most relevant to distinguish between phonological categories. Their ranking closely reflects the areas where there was more gestural activity (20%-40%-60%). But the RF shows that the 40% is the most relevant, then the 20% and 60%. These sections are the ones that are driving the main gestural differentiations, whereas the 80% functions as the anchor point for all the segments. The 100% shows that all the segments have a negative displacement, with /tʃ/ having the largest and /s/ the smallest displacement, and /t, ʃ/ with mid-values. In this case, we have evidence of distinctive patterns among segments, but also common patterns observed.

The Decision Tree analysis adds more crucial information on the stage at which these parameters are activated and the approximate length of displacement that drives the distinctions. These show again that the 40% is the most crucial area for gestural behaviour, but the 80% also triggers further distinctions. Results from this model also tells us the amount of displacement, and what direction, that is necessary to identify segment types. Combining all results, it shows that there is a cutoff at around 1.5mm, with alveolars being lower and palato-alveolars larger. In turn, fricatives have lower displacements than the stop counterparts. This higher displacement for stops can be understood as an articulatory requirement. Since these segments achieve complete constriction, it makes sense that they would require larger displacements of the tongue to reach its articulatory target.

In terms of the error for both models, stops show higher rates than fricatives. In addition, between /t/ and /tʃ/, the alveolar segment is the one that triggers the most errors. All errors are not random. For the palato-alveolars, in most of the errors they are confused by their counterpart with different manner of articulation. In the case of the alveolars, they are mainly confused with /ʃ/, and not the manner of articulation counterpart.

6. Limitations and Future Work

There are three main limitations identified in this paper. The first one is that the displacement measurements have been extracted for only female speakers. Further analysis would include male speakers to account for vocal tract size differences. Another limitation is that it has only been tested on voiceless coronal obstruents in English. We aim to test this with other segments, e.g. velars and palatals, as well as voiced vs voiceless classifications. Another limitation is that it was only trained on four consonants in a single vowel context, and it is yet to be shown whether it would perform well on a more diverse set of data. The final limitation is that the context for each token was fully controlled. Further investigation would reveal classification accuracy in cases where not all the tokens share the same context, especially in free speech.

7. Conclusions

In this paper, we have developed a Machine Learning Ensemble that analyses and classifies ToUS contours based on displacement measurements. Its high accuracy makes it a robust model to be implemented on new input data. Since it was trained on eight different speakers, it can be used to predict new data from different speakers. By analysing the variable of importance from the RF and the predictors at each node split in the Decision Tree, the approach allows having a clear understanding on what are the driving parameters when dealing with paired comparisons (e.g. alveolrars vs palato-alveolars (Place of Articulation), or stops vs fricatives (Manner of Articulation)). This model can be used to classify new data, given that it follows a similar format. The code and the model can be accessed through the following GitHub repository: https://github.com/simongonzalez/TongueUltrasoundAndML.

The methodological advancements presented here are relevant to the field of ToUS analysis in which identification and classification of articulatory landmarks can be automated, and thus, maximising the examination of the phonological implications of such differences.

8. Acknowledgements

I want to thank the anonymous reviewers for their comments and suggestions, which have improved this paper in many ways. The errors that remain are entirely my own responsibility.

9. References

[1] Wen, S., "Automatic Tongue Contour Segmentation using Deep Learning", Thesis, University of Ottawa, 2018.

[2] Tang, L., Hamarneh, G. and Bressmann, T., "A Machine Learning Approach to Tongue Motion Analysis in 2D Ultrasound Image Sequences", Machine Learning in Medical Imaging, 151-158, 2011.

[3] Stone, M., "A guide to analysing tongue motion from ultrasound images", Clinical Linguistics & Phonetics, 19(6-7):455-501, 2005.

[4] Gick, B., Campbell, F. and Oh, S., "A cross-linguistic study of articulatory timing in liquids", The Journal of the Acoustical Society of America, 110(5), 2001.

[5] Mielke, J., "An ultrasound study of Canadian French rhotic vowels with polar smoothing spline comparisons", The Journal of the Acoustical Society of America, 137(5), 2015.

[6] Roxburgh, Z., Cleland, J., Scobbie, J.M. and Wood, S.E., "Quantifying changes in ultrasound tongue-shape pre- and post-intervention in speakers with submucous cleft palate: an illustrative case study", Clinical Linguistics & Phonetics, 36:2-3, 146-164, 2022.

[7] Karimi, E., Ménard, L. and Laporte, C., "Fully-automated tongue detection in ultrasound images", Computers in biology and medicine, 111, 103335, 2019.

[8] Roon, K.D., Chen, W.-R., Iwasaki, R., Kang, J., Kim, B., Shejaeya, G., Tiede, M.K. and Whalen, D.H., "Comparison of auto-contouring and hand-contouring of ultrasound images of the tongue surface", Clinical Linguistics & Phonetics, 2022.

[9] Csapó, T.G. and Lulich, S.M., "Error analysis of extracted tongue contours from 2d ultrasound images", INTERSPEECH 2015, September 6-10, Dresden, Germany, 2015.

[10] Mozaffari, M.H. and Lee, W., "Deep Learning for Automatic Tracking of Tongue Surface in Real-time Ultrasound Videos, Landmarks instead of Contours", IEEE International Conference on Bioinformatics and Biomedicine (BIBM), 2785-2792, 2020.

[11] Jaumard-Hakoun, A., Xu, K., Roussel-Ragot, P., Dreyfus, G. and Denby, B., "Tongue contour extraction from ultrasound images based on deep neural network", ArXiv, abs/1605.05912, 2016.

[12] Fasel, I.R. and Berry, J., "Deep Belief Networks for Real-Time Extraction of Tongue Contours from Ultrasound During Speech", 20th International Conference on Pattern Recognition, 1493-1496, 2010.

[13] Zhu, J., Styler, W. and Calloway, I.C., "A CNN-based tool for automatic tongue contour tracking in ultrasound images", ArXiv, abs/1907.10210, 2019.

[14] Xu, K., Csap'o, T.G. and Feng, M., "Convolutional Neural Network-Based Age Estimation Using B-Mode Ultrasound Tongue Image", ArXiv, abs/2101.11245, 2021.

[15] Barros, F., Valente, A.R., Albuquerque, L., Silva, S.S., Teixeira, A.J. and Oliveira, C., "Contributions to a Quantitative Unsupervised Processing and Analysis of Tongue in Ultrasound Images", ICIAR 2020: Image Analysis and Recognition, 170-181, 2020.

[16] Gonzalez, S., "Gridlines approach for dynamic analysis in speech ultrasound data: A multimodal app", Journal of the Association for Laboratory Phonology 12(1):16,1-28, 2021.

[17] Davidson, L., "Comparing tongue shapes from ultrasound imaging using smoothing spline analysis of variance", Journal of the Acoustical Society of America, 120(1):407-15, 2006.

[18] Gu, C., "Smoothing spline ANOVA models", New York, NY: Springer, 2002.

[19] Chiu, C., Wei, P.C., Noguchi, M. and Yamane, N., "Sibilant Fricative Merging in Taiwan Mandarin: An Investigation of Tongue Postures using Ultrasound Imaging", Lang Speech. 63(4):877-897, 2020.

[20] Zharkova, N., Hewlett, N. and Hardcastle, W.J., "Coarticulation as an Indicator of Speech Motor Control Development in Children: An Ultrasound Study", Motor Control, 15:118-140, 2011.

[21] Barbier, G., Perrier, P., Payan, Y., Tiede, M.K., Gerber, S., Perkell, J.S. and Ménard, L., "What anticipatory coarticulation in children tells us about speech motor control maturity", PLOS ONE 15, 2020.

[22] Ostry, D.J. Keller, E. and Parush, A., "Similarities in the control of the speech articulators and the limbs: Kinematics of tongue dorsum movement in speech", Journal of Experimental Psychology: Human Perception and Performance. 9, 622, 1983.

[23] Strycharczuk, P. and Scobbie, J. M., "Velocity measures in ultrasound data. Gestural timing of post-vocalic /l/ in English", In Proceedings of the 18th International Congress on Phonetic Sciences, 10 August - 14 August, Glasgow, U.K, 2015.

[24] Wrench, A. and Scobbie, J.M., "Very high frame rate ultrasound tongue imaging", 2011.

[25] Carignan, C., "Using ultrasound and nasalance to separate oral and nasal contributions to formant frequencies of nasalized vowels", The Journal of the Acoustical Society of America, 143(5), 2588, 2018.

[26] Hueber, T., Aversano, G., Chollet, G., Denby, B., Dreyfus, G., Oussar, Y., Roussel, P. and Stone, M., "Eigentongue feature extraction for an ultrasound-based silent speech interface", in IEEE International Conference on Acoustics, Speech and Signal Processing, 1245-1248. Honolulu, HI: Cascadilla Press, 2007.

[27] Carignan, C. and Mielke, J., "Extracting articulatory signals from lingual ultrasound video using principal component analysis", MS, 2014.

[28] Cox, F., "Vowel transcription systems: An Australian perspective", International Journal of Speech-Language Pathology, 10:327-333, 2008.

[29] Li, M., Kambhamettu, C. and Stone, M., "Automatic contour tracking in ultrasound images", Clinical Linguistics & Phonetics, 19(6-7):545-554, 2005.

[30] Boersma, P. & Weenink, D., "Praat: doing phonetics by computer [Computer program]", Version 5.4.07, retrieved 22 March 2015 from http://www.praat.org/, 2005.

[31] Wrench, A., "Articulate Assistant Advanced User Guide", Edinburgh: Articulate Instruments Ltd., 2012.

[32] Liker, M., Zorić, A. V., Zharkova, N. and Gibbon, F. E., "Ultrasound Analysis of Postalveolar and Palatal Affricates in Croatian: A Case of Neutralisation", in S. Calhoun, P. Escudero, M. Tabain and P. Warren [Eds], Proceedings of the 19th International Congress of Phonetic Sciences, 3666-3670, Melbourne, Australia: Australasian Speech Science and Technology Association Inc., 2019.

[33] Micucci, M. and Iula, A., "Recent Advances in Machine Learning Applied to Ultrasound Imaging", Electronics 11(11):1800, 2022.

[34] 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.

[35] Breiman, L., "Random forests", Machine Learning, 45:5-32, 2001.

[36] Gareth, J., Witten, D., Hastie, T. and Tibshirani, R., "An introduction to statistical learning: with applications in R", New York:Springer, 2013.

[37] Kim, B., Tiede, M.K. and Whalen, D.H., "Evidence for pivots in tongue movement for dipthongs", in S. Calhoun, P. Escudero, M. Tabain and P. Warren [Eds], Proceedings of the 19th International Congress of Phonetic Sciences, 3666-3670, Melbourne, Australia: Australasian Speech Science and Technology Association Inc., 2019.

Training forced aligners on (mis)matched data: the effect of dialect and age

Tünde Szalay[1,2], Mostafa Shahin[2], Kirrie Ballard[1], Beena Ahmed[2]

[1]The University of Sydney, Sydney, Australia
[2]The University of New South Wales, Sydney, Australia
tuende.szalay@sydney.edu.au

Abstract

Training forced phonemic aligners for novel language varieties is non-trivial, as it requires aligned corpora. However, aligning novel corpora requires accurate forced aligners. To align AusKidTalk, an audio corpus of Australian English (AusE) speaking children, we trained three custom aligners on different datasets: age-matched American English (AmE) children, dialect-matched AusE-speaking adults, and their combination. Forced aligner performance using the three custom aligners and the Munich Automatic Segmentation System (MAUS) was evaluated against manual segmentation. The dialect-matched and combined custom aligners outperform MAUS, but the age-matched aligner does not. Our aligners' improved forced segmentation will reduce the time-need of manual correction.

Index Terms: forced phonemic alignment, accent differences, developmental differences, custom aligner for AusE-speaking children, audio corpus

1. Introduction

Segmenting acoustic data into phonemes is necessary for phoneme-level acoustic analysis in linguistics [1, 2]. While manual segmentation is considered the "gold standard" in terms of accuracy [2, 3], the use of forced aligners is recommended as manual phonemic alignment may take 800-times more than the length of the audio [4, 5]. Therefore, aligning large datasets is only possible using automatic forced aligners due to the time associated with manual alignment [4, 5].

During forced alignment, the orthographic transcription of the data is converted to phonemic transcription using a grapheme-to-phoneme pronunciation dictionary, and the phonemic transcription is mapped onto the acoustic data using acoustic models [2]. The acoustic model is created by pre-training the aligner using existing time-aligned speech corpora and a pronunciation dictionary, providing grapheme-to-phoneme mappings [2, 6]. Pre-trained models are not affected by speaker-specific idiosyncrasies, as they are trained on a large number of speakers, enabling them to generalise across them by learning speaker-independent characteristics of a language [1, 6, 7].

The performance of forced aligners is negatively impacted by domain- or population level linguistic differences between training- and novel data, such as differences between read and spontaneous speech, or between dialects [6, 7, 8, 9, 10]. Forced aligners trained on American English (AmE), while generally accurate on other dialects of English, produce larger errors in vowel boundaries as differences between training and novel data increase [8, 9]. For instance, an aligner trained on AmE, places 90% of automatic boundaries within 20 ms of manual boundaries for Received Pronunciation (RP), but only 75% for the Westray variety of Scots, a variety that shows larger differences from AmE than RP [8]. In Trinidad English, automatic vowel boundaries are overall accurate with 9–24 ms discrepancies from human alignment; however, Trinidad English-specific vowels show larger discrepancies [9]. Even small differences between automatic and human boundaries can have a roll-on effect, as vowel duration measured using forced and manual alignment may differ by up to 17 to 67 ms in Trinidad English, with Trinidad English-specific vowels showing the largest measurement differences [9]. Due to the adverse effect of accent differences, accent-specific aligners were developed for American, British, Australian, and New Zealand English [11].

Developmental differences between adults' and children's speech also have a negative effect on the accuracy of forced aligners [10]. As most aligners are pre-trained on adult speech, they show low accuracy on children's speech, with 69% to 79% agreement with human annotation [10]. Forced-aligner accuracy improves for older children compared to younger children, as children become more adult-like, and thus more accurately aligned [10]. Despite the adverse impact of age, no age-specific forced aligner has been pre-trained due to the lack of sufficiently sized and segmented children's corpora [12].

AusKidTalk, a large-scale corpus of Australian English (AusE) speaking children, is currently being developed to provide a corpus large enough for developing automated speech analysis tools for the novel variety of AusE-speaking children [12]. Developing a novel forced aligner for AusE-speaking children requires annotated training data, therefore AusKidTalk must, at least partially, be annotated using existing tools. However, AusKidTalk differs from the training data used in most available forced aligners in its accent (AusE vs. AmE) and age (children vs. adults). Therefore, we developed and tested three custom aligners with different pronunciation dictionaries and acoustic models, each trained on partially matching datasets for AusE-speaking children. The first acoustic model was trained on AmE-speaking children, thus training data matched target age, but not dialect. The second acoustic model was trained on AusE-speaking adults, thus training data matched target dialect, but not age. The third acoustic model was trained on both datasets, thus training data partially matched age and dialect. We evaluated the performance of our custom aligners by comparing it to human ground truth annotation as well as to the Munich Automatic Segmentation System (MAUS) [11].

2. Methods

2.1. Custom aligners

We developed three custom aligners, each with different acoustic models and pronunciation dictionaries (Fig. 1). The acoustic models were implemented using a Factored Time-Delay Neural Network in the Kaldi toolkit [13, 14], and trained on three, partially domain-matched datasets: AmE-speaking children (AmE

Child), AusE-speaking adults (AusE Adult), and on the combination of the two sets (Combined) (Fig. 1). The AmE Child model was trained on four children corpora yielding a total of 400 hours including single words and continuous speech – the Oregon Graduate Institute kids' speech corpus [15], the Carnegie Mellon University kids' speech corpus [16], the Colorado University Kids' corpus [17, 18], and the My Science Tutor Children's speech corpus [19]. Grapheme-to-phoneme conversion was provided by The Carnegie Mellon University (CMU) Pronouncing Dictionary [20]. The AusE adult model was trained on 800 hours of speech using the scripted, single word and continuous speech production tasks from the AusTalk corpus [21]. Grapheme-to-phoneme transcription was provided by orthographic- and phonemic transcriptions of the tasks used to elicit speech in AusTalk. The Combined model was trained on 1200 hours of speech using the AmE-speaking children's and the AusE-speaking adults' corpora. Grapheme-to-phoneme transcription was provided by the CMU Pronouncing Dictionary for the AmE-speaking children and by AusTalk transcriptions for the AusE-speaking adults. The Combined model used Multi-Task Learning to share consonant output layer between the dialects and had dialect-specific vowel output layers [22].

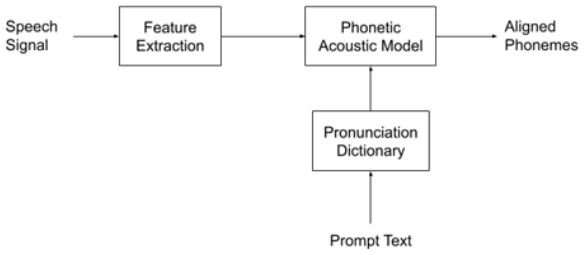


Figure 1: *Schematic diagram of forced aligner architecture.*

During forced alignment, the speech signal is divided into 25 msec frames with 15 msec overlap. Each frame is multiplied by a Hamming windowing function and 40 Mel-Frequency Cepstral Coefficients (MFCC) are extracted from each windowed frame (Fig. 1). The extracted features are fed into the acoustic model along with the expected phoneme sequence created through phonemic transcription of the prompts' orthographic transcription using the appropriate pronunciation dictionary – CMU Pronouncing Dictionary for the AmE Child model and corpus transcriptions for the AusE Adult model. The acoustic model assigns each frame to the most likely phoneme based on the pre-trained phoneme models and is constrained by the given phoneme sequence.

2.2. Test data

Forced-aligner performance was evaluated using data from the AusKidTalk corpus [12]. Speech recordings of eleven (M = 7, F = 4, aged from 4;10 to 11;11, mean = 7;7) children were extracted from the database. Children were native speakers of AusE without any speech disorders. Children produced 18 target words in a picture naming task (range = 11-15 words per child, mean = 13.9 words), giving a total of 153 words. Target words were extracted and saved as single-word wav files.

2.3. Forced and manual aligning

Words were force-aligned with our AmE Child, AusE Adult, and Combined custom aligners, using the sound files and their orthographic transcriptions. Words were force-aligned with the MAUS webtool, using the grapheme-to-phoneme (G2P) → MAUS pipeline without automatic speech recognition, and with the AusE pronunciation dictionary [11, 23, 24, 25]. Expert options were set to default; no custom rules were added. Sound files paired with matching text files containing orthographic transcription of the word with standard English spelling were uploaded to MAUS. MAUS was accessed on 27 August 2021, as well as on 06 June 2022. Results from 2022 are reported.

Manual segmentation was carried out by a trained phonetician in Praat [26], to provide ground truth segmentation prior to observing automatic alignment. Manual segment boundary placement was informed by periodicity, amplitude, and formant structure as presented in the waveform and the spectrogram.

All phoneme-level segmentation was carried out on wav files containing single words to prevent errors caused by confusion between words, such as mistaking the last segment of *snake* with the first segment of *key*, and to minimise the size of alignment errors. All aligners returned the results in Praat textgrids [26]. Boundary locations for all aligners (forced and manual) were extracted from the textgrids using Praat [26].

2.4. Analysis

A total of 816 (phoneme boundaries) × 4 (forced aligners) = 3,264 automatic boundaries were compared to manually placed boundaries. Boundary displacement between automatic and manual boundaries was calculated as the absolute value difference of manual minus automatic boundary [2]. Accuracy of automatically placed boundaries was calculated based on displacement: automatic boundaries were rated as correct when the boundary displacement was 20 ms or below, and as incorrect when displacement exceeded 20 ms [3]. Overlap rate between automatically segmented phonemes was calculated relative to the human annotation, using the time shared between human annotator and forced aligner (*Dur Shared*), the duration of the human aligner (*Dur Hum*), and the duration of the forced aligner (*Dur Forced*) using Equation 1 [2, 27].

$$Overlap = \frac{DurShared}{DurHum + DurForced - DurShared} \quad (1)$$

Equation 1 gives a score from 0 (representing no overlap) to 1 (representing complete overlap) for every phoneme. The distribution of Overlap rate was left-skewed, and bound from 0 to 1, making it conditionally beta-distributed [29]. As 0 and 1, despite being genuine outcomes of Overlap rate, cannot be included in the beta distribution (bound between 0 and 1, non-inclusive), Overlap rate was transformed to beta distribution using the weighted average (N Boundary = 3,264) and a constant 0.5 (Equation 2) [28, 29].

$$OverlapBeta = \frac{Overlap \times (NBoundary - 1) + 0.5}{NBoundary} \quad (2)$$

We constructed a generalised linear mixed effect model (GLM) with the dependent variable Accuracy (binomial family). As the 20ms threshold for accurate boundaries can indicate a quite large discrepancy, especially at fast speech rates, we constructed two more GLMs, one with Displacement (Gaussian family), and one with Overlap (Beta-transformed, Beta family) as dependent variables. The independent variable was Aligner (contrast coded, MAUS as baseline); Speaker was random intercept [30, 31]. p-values were calculated using Satterthwaite's

degrees of freedom method [32]. Planned comparisons with Bonferroni correction were used to compare the AmE Child, AusE Adult, and Combined custom aligners to each other [33]. All data analysis was done in R [34].

3. Results

The Combined custom aligner ($\beta = 0.325, z_{0.103} = 3.165, p = 0.0016$) and the AusE Adult custom aligner ($\beta = 0.254, z_{0.102} = 2.492, p = 0.0127$) produced significantly more accurate boundaries compared to MAUS. Accuracy decreased significantly when using the GenAm Child custom aligner compared to using MAUS ($\beta = -0.489, z_{0.1} = -4.896, p < 0.0001$) (Fig. 2).

Planned comparison confirmed that MAUS is significantly less accurate than the Combined ($p = 0.0093$), and more accurate than the GenAm Child custom aligner ($p < 0.0001$). Contrary to our GLM model, planned comparison did not show a significant difference between MAUS and the AusE Adult custom aligner ($p = 0.0761$). Planned comparison revealed that the GenAm Child custom aligner performs significantly less accurately than the Combined ($p < 0.0001$) and the AusE Adult ($p < 0.0001$) custom aligners. No difference was found between the Combined and the AusE Adult aligners ($p = 1$).

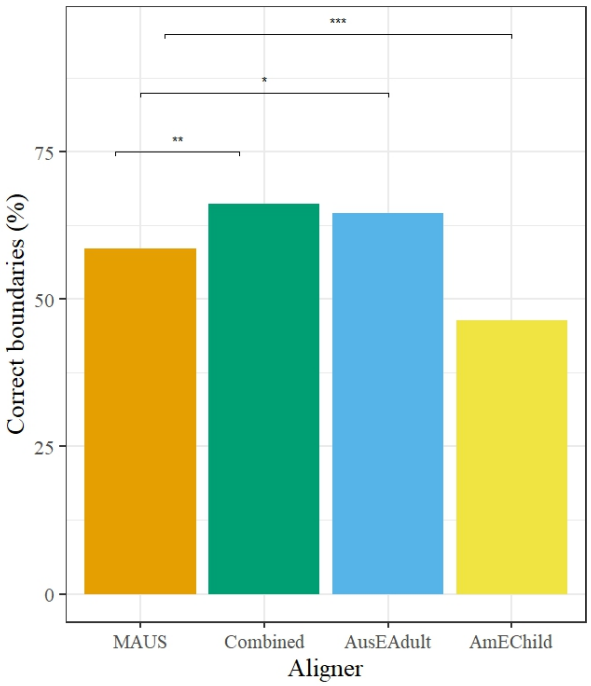


Figure 2: *Boundary accuracy. Significance taken from GLM.*

Choice of forced aligner had no significant effect on boundary displacement in the GLM (Fig. 4) or in the planned comparisons. Boundary displacement was non-significantly smaller (i.e., better) using the AusE Adult ($\beta = -3.452$) and the AmE Child aligners ($\beta = -0.441$) and non-significantly larger (i.e., worse) using the Combined aligner ($\beta = 3.892$) than MAUS.

Overlap rate increased significantly using the custom forced aligners compared to MAUS (Combined: $\beta = 0.196, z_{0.048} = 4.112, p < 0.0001$; AusE Adult: $\beta = 0.268, z_{0.048} = 5.604, p < 0.0001$; AmE Child: $\beta = 0.104, z_{0.048} = 2.155, p < 0.0311$). Planned comparisons confirmed the significantly larger (i.e., better) overlap rate for the Combined ($p = 0.0002$) and the AusE Adult ($p < 0.0001$) aligners compared to MAUS. Contrary to the GLM results, planned comparison showed no difference between MAUS and the AmE Child custom aligner ($p = 0.1873$). Planned comparison revealed that the AmE Child custom aligner shows significantly less overlap with the human annotation compared to the AusE Adult ($p = 0.0043$) aligner but not from the Combined ($p = 0.3396$) aligner. No difference was found between the Combined and the AusE Adult aligners ($p = 0.8045$).

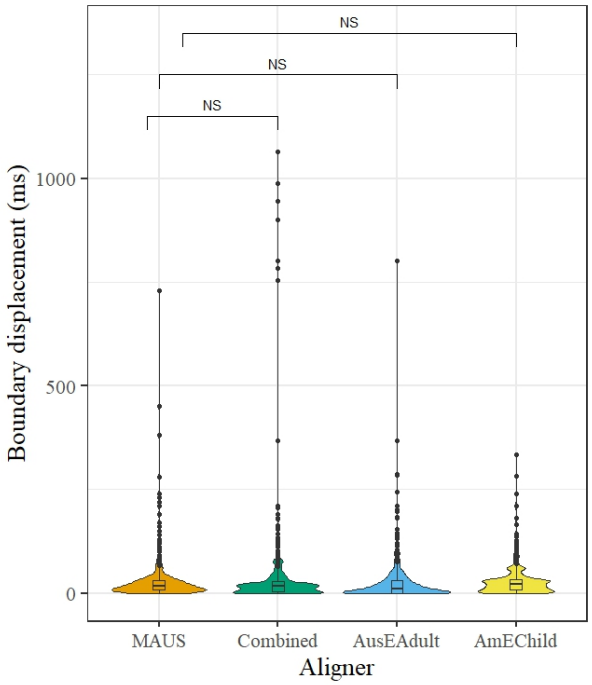


Figure 3: *Boundary displacement. Significance taken from GLM.*

4. Discussion

Our goal was to explore the effect of age- and dialect mismatch on forced-aligning AusE-speaking children's speech. The age-matched, but dialect mismatched custom aligner used an acoustic model trained on AmE-speaking children and a North American pronunciation dictionary; the age-mismatched but dialect-matched aligner used an acoustic model trained on AusE-speaking adults and an AusE pronunciation dictionary; the Combined acoustic model used both datasets with both pronunciation dictionaries. Our AusE Adult and Combined custom aligners outperformed MAUS, both of which used accent-matched training data and pronunciation dictionary. However, performance decreased when our custom aligner used an acoustic model trained on age-matched data and an accent-mismatched pronunciation dictionary. Overall quality of all forced aligners remained low (Table 1), therefore, manual correction of automatically placed boundaries is required for phoneme-level linguistic analysis of AusE-speaking children.

4.1. Custom aligners: the effects of dialect and age

Using the GenAm Child forced-aligner for AusE-speaking children with an age-matched but dialect mismatched acoustic

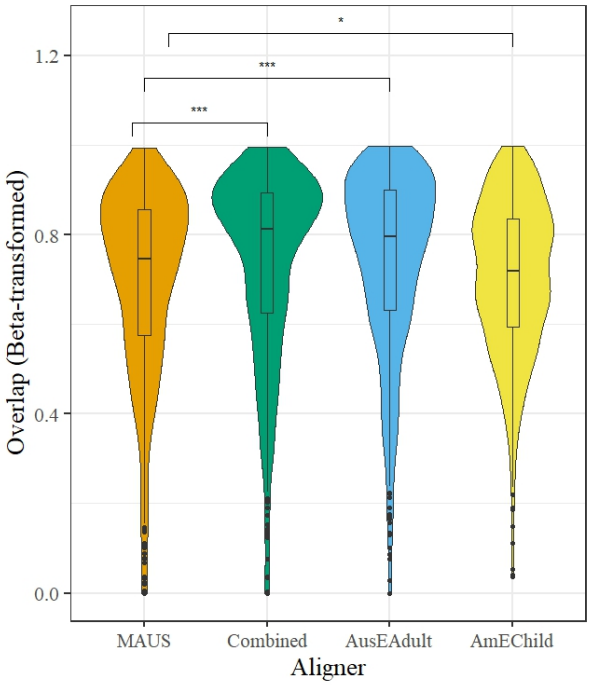


Figure 4: *Overlap rate. Significance taken from GLM.*

Table 1: *Summary of results with accuracy (Acc., %), mean displacement (Disp., ms), and mean overlap rate (Overlap, 0-1, inclusive) for each forced aligner.*

Aligner	Age	Dialect	Acc. (%)	Disp. (ms)	Overlap (0-1)
MAUS	×	✓	59	28	0.69
AusE Adult	×	✓	65	24	0.74
AmE Child	✓	×	46	27	0.71
Combined	✓	✓	66	32	0.73

model yields worse performance than using a similar custom forced aligner with accent-matched acoustic models. The poor performance of the AmE Child aligner indicates that accent differences (AmE vs. AusE) outweigh developmental differences (children vs. adults), as accent differences between AmE and AusE adversely impact the usability of the acoustic model trained on AmE data and of the pronunciation dictionary.

The AmE Child aligner's acoustic model might perform poorly due to dialectal differences. For instance, acoustic vowel differences between AmE training and AusE test data (e.g., *goose* contains back /uː/ in GenAm, but central /ʉː/ in AusE) might lead to incorrect feature mapping between the raw speech and the phonemes. In addition, similarities between the AmE Child training and the AusE child test data might be smaller than expected. For instance, patterns for acquiring /l/, a late acquired sound in both dialects, are similar but not identical [37]. Differences in developmental trajectories between AmE and AusE may cause incorrect feature-to-phoneme mapping and further reduce the suitability of the acoustic model. Thus, using accent-matched AusE training data for AusE-speaking children is required due to the considerable accent differences between GenAm and AusE and the small age-related similarities between children of the two dialects.

The AmE Child aligner's errors in words containing a post-vocalic /ɹ/ in AmE, but not in AusE (e.g., *car, spiderweb*) can be attributed to applying AmE grapheme-to-phoneme mapping onto AusE speech. The pronunciation dictionary in the AmE Child acoustic model maps the letter "r" onto the phoneme /ɹ/ in word-final and pre-consonantal positions, as AmE is a rhotic accent, allowing /ɹ/ in word-final, pre-consonantal, and pre-vocalic positions [36]. In contrast, AusE is a non-rhotic accent, in which /ɹ/ only occurs pre-vocalically (e.g., /ɹ/ occurs in *red*, but not in *car* and *spiderweb*) [35]. As a result, the GenAm Child aligner attempts to map the single AusE vowel into a vowel-/ɹ/sequence, resulting in a vowel offset placed before the acoustic end of the vowel and an unnecessary /ɹ/ interval. Errors caused by /ɹ/-insertion show the detrimental effect of incorrect grapheme-to-phoneme mapping, although previously no such effect of mismatched pronunciation dictionary was found [2]. Therefore, using a dialect-matched, AusE pronunciation dictionary is recommended.

Combining the training data for AusE-speaking adults with GenAm-speaking children increased the data-need of our custom aligner, without leading to a significant improvement in performance. A non-significant reduction in the number of errors, coupled with an increase in the size of errors was observed. As the time needed for manual correction of automatic segmentation depends on the number of errors rather than on the size of errors, even a small difference in accuracy is likely to lead to a considerable reduction in the time and resources required for manual correction. As both accent-matched and age-matched acoustic data are readily available through open-source corpora [15, 16, 17, 18, 19, 21], combining accent-matched and age-matched training data is recommended.

4.2. MAUS and the custom aligners

Our accent-matched aligners outperformed MAUS. The acoustic models of MAUS and our custom aligner's AusE Adult and Combined acoustic models were trained using the same AusTalk dataset [38]. However, MAUS uses the Hidden Markov Toolkit, whereas our custom aligner uses the Factored Time-Delay Neural Network in the Kaldi toolkit [14, 39]. The improved performance of our custom aligner is attributed to the more advanced network used by Kaldi. Similarly, the Kaldi-based Montreal Forced Aligner outperformed other aligners with the Hidden Markov Toolkit [2].

MAUS was accessed on 27 August 2021 and on 06 June 2022. Between 2021 and 2022, the AusE dictionary for MAUS was corrected, and the grapheme-to-phoneme, syllable, and word stress models were re-trained. Using MAUS 2021 versus MAUS 2022 with the same settings did not change the results, despite some improvements: from 2021 to 2022, accuracy of MAUS improved from 58% to 59%, displacement from 28 ms to 27 ms, while overlap rate remained the same (0.69). MAUS allows a high-level of customisation in forced aligning, including the addition of custom rules. These custom features were not used in our current study. It is possible that custom settings would improve forced alignments.

5. Conclusion and future directions

To date, the best-performing forced aligner is our custom built forced aligner using a combined acoustic model of AusE-speaking adults and AmE-speaking children. The main drawback of our custom aligner is its lack of accessibility - while MAUS is easily accessible through its web interface, our aligner is located on a private server. Therefore, future work will include sharing our custom built aligner.

6. Acknowledgements

This project was supported by the Australian Research Council LE190100187 grant. We would like to thank the participants who provided their voice for the AusKidTalk project, and without whom this research would not have been possible.

7. References

[1] Fromont, R. and Watson, K., "Factors influencing automatic segmental alignment of sociophonetic corpora", Corpora, 11(3):401—431, 2016.

[2] González, S., Grama, J., and Travis, C., "Comparing the accuracy of forced-aligners for sociolinguistic research", Linguistics Vanguard, 6(1).

[3] Cosi, P., Falavigna, D., and Omologo. M., "A preliminary statistical evaluation of manual and automatic segmentation discrepancies", Proc. Eurospeech, 693–696, 1991.

[4] Gibbon, D., Moore, R., and Winski, R., Handbook of standards and resources for spoken language systems, 1997.

[5] Schiel, F., Draxler, C., Baumann, A., Ellbogen, T., and Steffen, A. The production of speech corpora, 2012.

[6] Brognaux, S., Roekhaut, S., Drugman, T., and Beaufort, R., "Automatic phone alignment", Proc Int Conf on NLP., 300–311.

[7] Chen, L., Liu, Y., Harper, M. P., Maia, E., and McRoy, S,., "Evaluating Factors Impacting the Accuracy of Forced Alignments in a Multimodal Corpus", Proc LREC, 2004.

[8] MacKenzie, L., and Turton, D., "Assessing the accuracy of existing forced alignment software on varieties of British English", Linguistics Vanguard, 6(s1), 2020.

[9] Meer, P., "Automatic alignment for New Englishes: Applying state-of-the-art aligners to Trinidadian English." JASA 147(4):2283–2294, 2020.

[10] Mahr, T. J., Berisha, V., Kawabata, K., Liss, J., and Hustad, K. C., "Performance of forced-alignment algorithms on children's speech", J. Speech Lang. Hear. Res. 64(6S):2213-2222 (2020).

[11] Kisler, T.,Reichel, U. D., and Florian Schiel "Multilingual processing of speech via web services" Computer Speech & Language, 45:326–347, 2017.

[12] Ahmed, B., Ballard, K., Burnham, D., Tharmakulasingam S., Mehmood, H., Estival, D., Baker, E., Cox, F., Arciuli, J., and Benders, T., "AusKidTalk: An Auditory-Visual Corpus of 3-to 12-year-old Australian Children's Speech", ISCA, 2021.

[13] Povey, D., Cheng, G., Wang, Y., Li, K., Xu, H., Yarmohammadi, M., and Khudanpur, S., "Semi-orthogonal low-rank matrix factorization for deep neural networks", Proc Interspeech, 3743-3747, 2018.

[14] Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P. and Silovsky, J., "The Kaldi speech recognition toolkit", Proc IEEE workshop on automatic speech recognition and understanding, 2011

[15] Shobaki, K., Hosom, J.-P., and Cole, R. A., "The OGI kids' speech corpus and recognizers", Proc Sixth Int Conf on Spoken Language Processing, 2000.

[16] Eskenazi, M., Mostow, J., and Graff, D., "The CMU Kids Corpus LDC97S63", LDC database, 1997.

[17] Cole, R., Hosom, P., and Pellom, B., "University of Colorado prompted and read children's speech corpus", Technical Report TR-CSLR-2006-02, 2006.

[18] Cole, R. and Pellom, B., "University of Colorado read and summarized story corpus", Technical Report TR-CSLR-2006-03, 2006.

[19] Ward,W., Cole, R., and Pradhan, S., "My Science Tutor and the MyST Corpus," 2019.

[20] Weide, R., The Carnegie Mellon pronouncing dictionary, 1998.

[21] Burnham, D., Estival, D., Fazio, S., Viethen, J., Cox, F., Dale, R., Cassidy, S., Epps, J., Togneri, R., Wagner, M., Kinoshita, Y., Göcke, R., Arciuli, J., Onslow. M., Lewis, T., Butcher, A., and Hajek, J., "Building an audio-visual corpus of Australian English: large corpus collection with an economical portable and replicable Black Box", Proc Interspeech, 841–844, 2011.

[22] Caruana, R., "Multitask learning", Machine learning 28(1):41-75, 1997.

[23] Reichel, U. D. "PermA and Balloon: Tools for string alignment and text processing", Proc Interspeech, 2012.

[24] Schiel, F. "Automatic Phonetic Transcription of Non-Prompted Speech", Proc ICPhS 607-610, 1999.

[25] Schiel, F. "A Statistical Model for Predicting Pronunciation", Proc ICPhS, 2015.

[26] Boersma, P., and Weenink, D., "Praat: doing phonetics by computer". Version 6.1.41, retrieved 25 March 2021 from www.praat.org

[27] Paulo, S., and Oliveira, L. C. "Automatic phonetic alignment and its confidence measures", Proc. Advances in Natural Language Processing, 36–44, 2004.

[28] Macmillan, N.A., and Creelman, D.C., "Detection theory: a user's guide", Lawrence Erlbaum Associates, 2005.

[29] Smithson, M., and Verkuilen, J., "A better lemon squeezer? Maximum likelihood regression with beta-distributed dependent variables", Psychological methods 11(1), 54–71, 2006.

[30] Bates, D., Mächler, M., Bolker, B., and Walker, S. "Fitting Linear Mixed-Effects Models Using lme4", J Stat Softw 67(1):1–48, 2015.

[31] Brooks, M., Kristensen, K., van Benthem, K. J., Magnusson, A., Berg, C. W., Nielsen, A., Skaug, H. J., Maechler, M., and Bolker B. M., "glmmTB: Balances Speed and Flexibility Among Packages for Zero-inflated Generalized Linear Mixed Modeling", The R Journal 9(2):378–400, 2017.

[32] Kuznetsova, A., Brockhoff, P. B., and Christensen, R. H. B., "lmerTest Package: Tests in Linear Mixed Effects Models", J Stat Softw 82(13):1–26, 2017.

[33] Lenth, R.V., "emmeans: Estimated Marginal Means, aka Least-Squares Means", 2021.

[34] R Core Team, "R: A Language and Environment for Statistical Computing", R Foundation for Statistical Computing, 2021.

[35] Cox, F., and Fletcher, J., Australian English pronunciation and transcription, Cambridge University Press, 2017.

[36] Labov, W., Ash, S., and Boberg, C., The Atlas of North American English, Phonetics, Phonology and Sound Change, Gruyter Mouton, 2008.

[37] Lin, S., and Demuth, C., "Children's acquisition of English onset and coda /l/: Articulatory Evidence", J. Speech Lang. Hear. Res. 58:13–27, 2015.

[38] Cassidy, S., Estival, D., and Cox, F., "Case study: the AusTalk corpus", in N. Ide, and J. Pustejovsky, J. [Ed] Handbook of linguistic annotation, 1287-1301, Springer 2017.

[39] Young, S., Evermann, G., Hain, T., Kershaw, D., Xunying A. L., Odell, J., Ollason, D., Povey, D., Valtchev V., and Woodland, P., "The HTK Book (For Version 3.4)" Cambridge University Engineering Department, 2006.

A Comparison of Machine Learning Algorithms and Human Listeners in the Identification of Phonemic Contrasts

Paul Reid, Ksenia Gnevsheva, Hanna Suominen

Australian National University

reidp02@gmail.com, ksenia.gnevsheva@anu.edu.au, hanna.suominen@anu.edu.au

Abstract

To elucidate the processes by which automatic speech recognition (ASR) architectures reach transcription decisions, our study compared human and ASR responses to stimuli with manipulated cues for stop manner (burst, silence, and vocalic onset) and voicing (voice onset time, aspiration amplitude, and vocalic onset). Fourteen participants and two ASR systems completed a forced-response identification task. Results indicated that the cues were of perceptual significance for human participants, and though weighted differently, significant predictors of ASR output. This demonstrated that ASR systems may be relying on the same key acoustic information as do human listeners for phonemic classification.
Index Terms: applied linguistics, English, evaluation studies, natural language processing, speech perception, speech recognition software

1. Introduction

Although an increasing body of studies compare human speech perception and automatic speech recognition (ASR) as a performance evaluation metric in ASR [1], [2], [3], [4], acoustic cues remain potentially under-utilised as a tool to explore phonemic representations in ASR systems in comparison to humans. Human listeners use voice onset time (VOT), vowel first formant transitions, and aspiration amplitude as voicing cues for word-initial stops in English [5], [6], [7], [8], [9]. In stop manner perception, the burst of noise accompanying stop release, a rising first formant transition, and silence cue stop manner [10], [11], [12], [13]. Human listeners exhibit the phenomena of trading relationships [14], categorical perception [15], and cue hierarchies [9].

This study compared the significance of acoustic cues for stop consonant voicing and manner in human and ASR perception with the following research questions:

1) How are acoustic cues for stop consonant voicing and manner used by human listeners?

2) How are acoustic cues for stop consonant voicing and manner used by ASR systems?

3) What are the differences between ASR systems and human listeners in the use of these acoustic cues?

2. Method

2.1. Stimuli

2.1.1. Base stimuli

Base stimuli were created, which were then manipulated to alter the value of the acoustic cues under investigation using PRAAT [16]. The recording was conducted by a speaker of Australian English, using a Rode NT1A microphone, recorded at 44.1 kHz in .wav format on a computer. The carrier phrase 'I am going to say X now' was repeated 30 times for each of the three words 'pat', 'bat', and 'stay'. 25 representative tokens for each target were selected and then annotated and modified.

To create the 'pat'-'bat' base stimuli, the vocalic portion of the 25 representative 'bat' stimuli was spliced with a representative region of 80 ms of aspiration from one selected 'pat' stimulus. Each of the 25 base stimuli was annotated into two sections: aspiration (80 ms) and vowel onset (100 ms).

To generate the 'stay'-'say' base stimuli, (1) a representative burst was selected from one of the 'stay' recordings and was inserted into each of the other representative 'stay' recordings in place of their burst; (2) the same 200 ms of silence, extracted from the audio recording between stimuli, was inserted between the fricative and the burst for each stimulus. Each of the 25 'stay'-'say' base stimuli was annotated into three sections: silence (200 ms), burst (20 ms), and vowel onset (50 ms).

2.1.2. Human perception experiment stimuli

The final stimuli for the human perception experiment were generated in five groups from one selected 'pat'-'bat' base and in four groups from one selected 'stay'-'say' base.

The first three groups of 'pat'-'bat' stimuli had the vowel onset removed. The first group had the amplitude of the aspiration unchanged at 100%; the second and third groups had the amplitude of the aspiration increased to 200% and reduced to 50%, respectively. Within these three sets, the length of the VOT was then varied between 0 and 45 ms, in 3 ms increments, resulting in 48 stimuli (16 per group). The last two groups of 'pat'-'bat' stimuli had the aspiration amplitude left unmodified (i.e., 100%). The first of these groups had VOT reduced to 30 ms; the other had VOT reduced to 45 ms. For both these sets of stimuli, the vowel onset was removed in 10 ms increments from the left until all 100 ms was removed, which resulted in a total of 22 stimuli.

The first two groups of 'stay'-'say' stimuli had the entire vowel onset portion removed, and the second group also had the burst entirely removed. For both the with-burst and burst-less stimuli, the silence was varied between 0 and 200 ms in 10 ms increments resulting in 42 stimuli (21 per group). The final two groups of 'stay'-'say' stimuli first had the burst removed entirely, then the first of the two groups had silence set to 30 ms, the other had silence set to 100 ms. For both groups, the 50 ms of vowel onset was removed in 5 ms intervals from the left, resulting in 22 stimuli (11 per group).

2.1.3. ASR stimuli

Two sets of stimuli were processed by the ASR systems; the first allowed for a broad, multivariate analysis of the decision

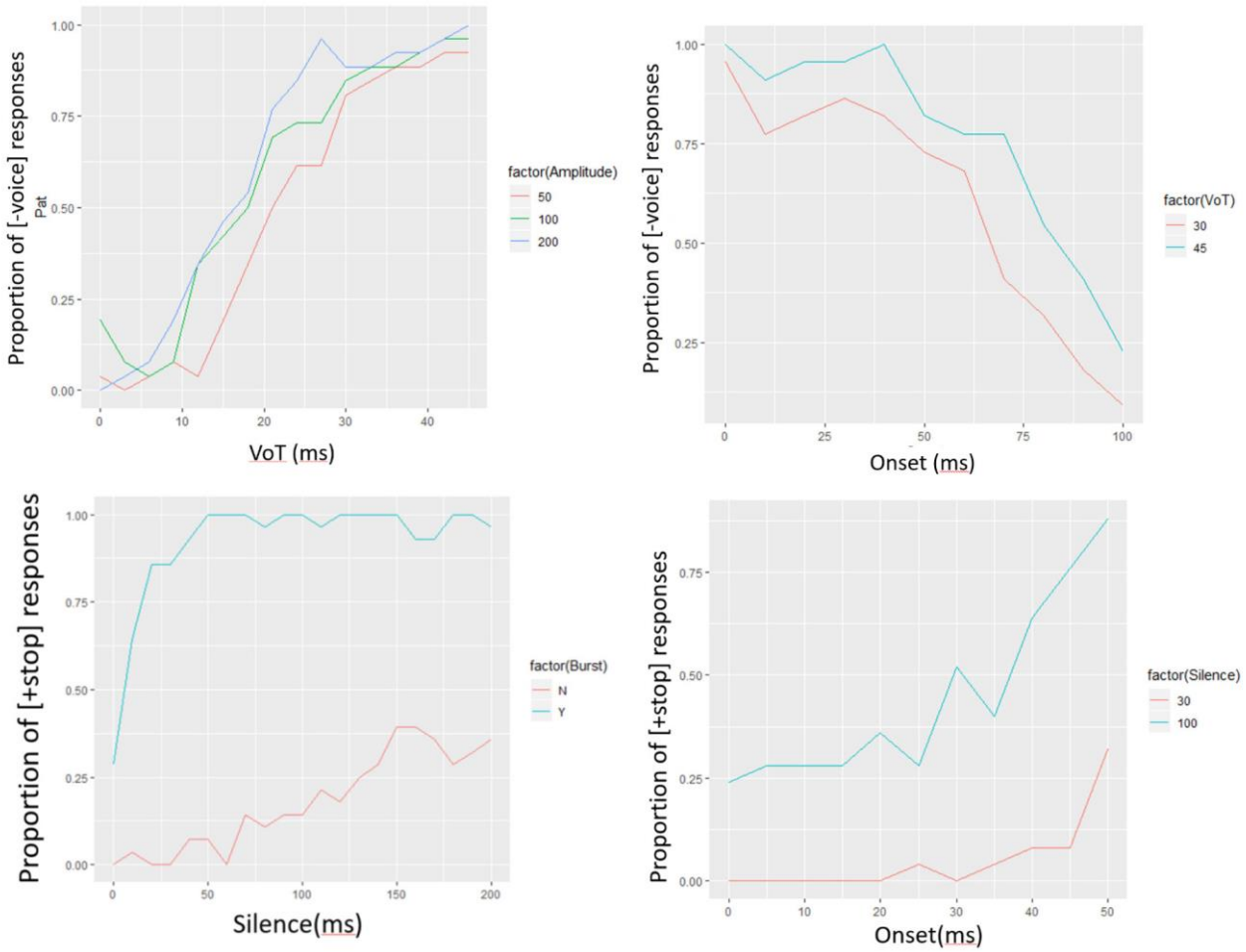


Figure 1: *Predictors of human participants' [-voice] and [+stop] responses.*

space of the ASR models (computer-only stimuli), whilst the second allowed a direct comparison between the ASR transcriptions and the responses in the human perception experiment (human comparison stimuli).

The computer-only 'pat'-'bat' stimuli were generated from each of the 25 base stimuli. The amplitude of the aspiration was varied between 50% and 200% of its original amplitude in 50% increments. For each of these sound files, the aspiration was cut in 10 ms increments to generate a new stimulus each time, starting from the right, until the aspiration was entirely removed. Finally, for each of those sound files, the vowel onset was cut in 10 ms increments from the left until the transition was entirely removed, which resulted in a total of 9900 stimuli.

The computer-only 'stay'-'say' stimuli were generated by modifying each of the 25 base stimuli. The vowel onset was cut in 5 ms increments, starting from the left, until the vowel onset section was entirely removed. For each of these, the burst was cut in 10 ms increments, starting from the right until the burst was entirely removed. Finally, the silence was reduced in 10 ms increments until entirely removed. This resulted in a total of 9075 individual 'stay'-'say' stimuli.

The acoustic manipulations for the human comparison stimuli were the same as the human perception experiment stimuli, carried out on all 25 base stimuli. This resulted in a total of 1,750 'pat'-'bat' stimuli and 1,600 'stay'-'say' stimuli.

2.2. Human perception experiment

The human perception experiment, which took approximately 30 minutes to complete, was conducted on a computer using the Qualtrics [17] interface. The participants were 14 monolingual speakers of Australian English between the ages of 20 and 40 (10 males and 4 females). The participants chose between 'pat' and 'bat' or 'stay' and 'say' after being presented with an audio stimulus of the form "I am going to say X now". The stimuli were placed into blocks by stimulus group, and the order of these blocks, the order of the individual stimuli within the blocks, and the order of the choices was randomized for each participant. This process was then repeated so that each block and stimulus was heard twice by each participant, which resulted in a total of 268 responses. Ethical approval (Protocol 2019/679) was obtained from the Human Research Ethics Committee of the Australian National University (ANU), and each study participant provided written informed consent.

2.3. ASR systems

The two ASR systems included in the experiments were Mozilla's DeepSpeech [18] and Google Speech-To-Text [19]. Mozilla's DeepSpeech is an open source, end-to-end ASR system based on the Recurrent Neural Network architecture. Version 0.6.0 pre-trained on American English was used for

ease of implementation. The default parameters were used when running the model, which was completed using a python script as per the documentation. Google's Speech-To-Text service is a cloud-based deep-learning based speech recognition system, which provides transcription services for 120 languages and dialects, including Australian English.

Non-target transcriptions represented no more than 3% of the total output for each set and were excluded from analysis. For each point in acoustic space represented by the values of the acoustic variables for the stimuli, the proportion of [-voice] and [+stop] output was calculated.

2.4. Analysis

Analysis of Variance (ANOVA) was conducted for both the human perception results and the output from the human comparison stimuli by the ASR systems. The independent variables tested were the acoustic variables modified for each group, as well as the interaction between them. Participant was used as an error term. Linear Discriminant Analysis (LDA) was used in the analysis of the computer-only output. An advisor from the ANU Statistical Consulting Unit was involved in this analysis design and its outcome reporting.

3. Results

3.1. Human listeners

Significant predictors of human listeners' ***voicing*** judgments in Group 1 stimuli were amplitude ($p < .001$) and VOT ($p < .001$): [-voice] responses increase with increase in VOT and amplitude (Figure 1, top left). Vowel onset ($p < .001$) and VOT ($p < .001$) were significant predictors of voicing in Group 2 stimuli: [-voice] responses decrease with longer onset and lower VOT (Figure 1, top right). Burst ($p < .001$), Silence ($p < .001$), and their interaction ($p < .05$) were significant predictors of ***manner of articulation*** judgments: in the with-burst condition [+stop] responses dominated, with a sharp drop at short silence; in the burst-less condition there was a slow, steady increase in the proportion of [+stop] responses (Figure 1, bottom left). Vowel Onset ($p < .001$), Silence ($p < .001$), and their interaction ($p < .001$) were significant predictors of [+stop] judgments: 100 ms silence was associated with a higher proportion of [+stop] responses, with a wide increase as vowel onset increased; [-stop] responses dominated at 30 ms silence with a slight increase at the longest onset values (Figure 1, bottom right).

3.2. ASR

The two ASR models demonstrated a distinct difference in ***voicing*** cues that were required to generate a [+stop] output, although this investigation was not the intent of the 'pat'-'bat' stimuli. DeepSpeech produced a very limited number of [+stop] transcriptions, instead outputting a fricative. Consequently, these results are excluded from analysis. For Google, as VOT increased, the likelihood of a [-voice] judgement also increased, except for VOT = 10 ms where [-voice] judgement dominated. As onset increased and amplitude decreased, so did the likelihood of a [+voice] judgement. LDA suggested that the three acoustic variables in combination were very powerful predictors of the model output. [+voice] and [-voice] were most widely separated by Vowel Onset and VOT, with the magnitude of the weighting for Amplitude (0.005) in Linear Discriminant 1 considerably lower than both Onset (0.02) and VOT (0.04).

The most obvious feature of the DeepSpeech [+stop] results related to the ***manner of articulation.*** Namely, the length of silence had a strong influence on the output, with an almost categorical [+stop] output for all values of Silence longer than 10 ms. As for Google, for Silence > 60 ms, [+stop] dominated regardless of the value of the other acoustic variables. The presence of the entire burst strongly cued [+stop]. At other Burst values, Onset remained a significant factor in determining the output until the limiting value of Silence was reached, which forced the [+stop] output.

LDA indicated that the three acoustic variables were strong predictors of the output for both models. For Google, LDA implied considerable separation for all three acoustic variables, whereas for DeepSpeech, Silence and Burst were well separated, whereas Onset much less so. For Google, LD1 weights of Silence (0.209), Onset (0.031), and Burst (0.017) were all relatively even while for DeepSpeech, Silence had the largest absolute LD1 weight (-0.053), followed by Burst (-0.01), and Onset (-0.003) was weighted very lightly.

3.3. Human vs ASR comparison

For Google, only the simple effect of VOT ($p < .001$) was shown to be significant for the human comparison stimuli for Group 1: low values of VOT are associated with predominantly [+voice] output (Figure 2, top left). VOT ($p < .001$) and Onset ($p < .001$), as well as their interaction, were significant for Group 2: [-voice] dominated at low values of Onset and decreased sharply at high values of Onset, with a higher value associated with a longer VOT (Figure 2, top right). The effect of Burst ($p < .001$) and the Silence:Burst interaction ($p < .001$) were significant for Group 3: in the with-burst condition, [+stop] was cued the majority of the time; in the burst-less condition, there was a steep rise from [-stop] to [+stop] output (Figure 2, bottom left). Silence ($p < .001$), Onset ($p < .001$), and the Silence:Onset interaction ($p < .001$) were significant for Group 4: longer silence resulted in categorical [+stop] responses while with shorter silence [+stop] increased as the vowel onset increased (Figure 2, bottom right).

For DeepSpeech, the simple effects of Burst ($p < .001$), Silence ($p < .001$) and the Burst:Silence interaction ($p < .001$) were significant for Group 3: [+stop] responses increased sharply at low silence duration, with a later rise in the burst-less condition. Onset ($p < .001$), Silence ($p < .001$) and the Onset:Silence interaction ($p < .001$) were significant for Group 4: [+stop] dominated for both Silence conditions, with a slight decrease in the long Silence condition at low Onset values.

4. Discussion

In human listeners, the significance of VOT [8], [9], vowel onset [5], [7], [8], and aspiration amplitude [9] in making stop voicing judgements in word initial position; and silence [10], [12], vowel onset [11], [13], and burst [13], [16] in making stop manner judgements in word medial position were confirmed. Our results supported claims that categorical effects are present in human perception of stop manner [20], [21] and stop voicing [22] contrasts. Previous research findings demonstrating a trading relationship between VOT and aspiration amplitude [9], and VOT and first formant [8] for stop voicing, and burst and silence [16] for stop manner were further supported. Finally, certain acoustic cues were only important within ambiguous values of more dominant cues, which supported the existence of a hierarchical relationship between the acoustic cues that were perceptually significant for a given phonemic contrast [9].

In ASR systems, for the stop manner contrast investigated, Burst, Onset, and Silence were all significant predictors for

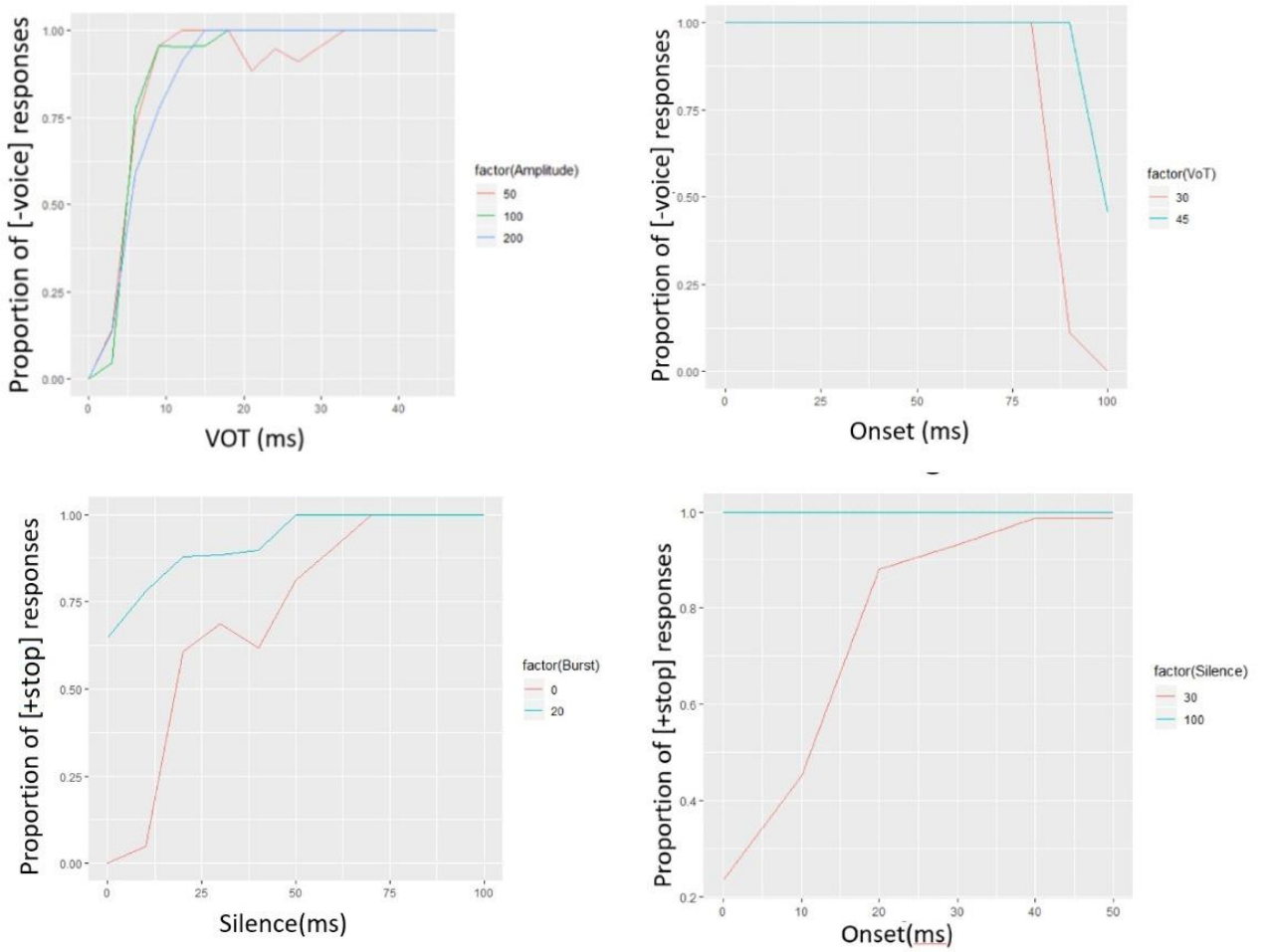


Figure 2: *Predictors of Google's ASR system responses.*

both systems; VOT, aspiration amplitude, and Onset were also significant predictors of stop voicing for Google. Further, a hierarchical relationship between the cues associated with a particular phonemic contrast and trading relationships between those acoustic cues were also present in ASR responses, in parallel with the phenomena observed in the human participants. The sharp gradient of the decision boundaries in response to changes in acoustic variables in stimuli also mirrored the categorical effect in humans.

The presence of these phenomena in both humans and ASR systems provided evidence to suggest that learned phonemic representations in the ASR systems may be based around the same kinds of acoustic information as humans use as their learning mechanisms. However, both experimented ASR models incompletely replicated the way in which humans make phonemic decisions, implying that the parallel between ASR and humans is not direct. In general, decision boundaries of the Google model appeared to more successfully replicate those from human perceptual results than DeepSpeech. Therefore, our results supported the claim that the sophistication of the phonemic representations of ASR systems is highly variable depending on the architecture used [23]. This finding highlighted the complexity of learning accomplished by ASR for making phonological judgements.

High-performance ASR systems, in particular neural network-based systems, underpinned by their methodological paradigm shift since 2010s explained by advancing deep learning, are capable of generating sophisticated representations of natural phonological classes [24], [25]. Since both ASR models, to varying degrees, used the same set of acoustic cues as humans to make phonemic decisions, this also implied that using these cues may be a natural consequence of optimising separation between categories based on the distribution of acoustic variables in that language. This was especially clear in the case of DeepSpeech, given its end-to-end architecture, since there was essentially no explicit acoustic engineering [26], and any phonetic cues learned to assist in the phonemic classification were based purely on distributional information within those phonemic categories.

If ASR systems develop comparable phonemic representations based on the same acoustic variables as humans, the theoretical ramifications could be significant. It may provide evidence for an auditory speech perception such as the Auditory Model [27], as it is unclear how an ASR system would represent phonemes in terms of articulatory gestures, as posited by Motor Theory [28]. It could also provide evidence for learning models based on the distributional properties of speech, such as statistical learning model [29]. It also presents a novel method for evaluating ASR performance by benchmarking against human perception; however, one must account for the spectral and temporal redundancies of acoustic cues [30] and covariance of acoustic cues in natural speech [31].

5. Acknowledgements

We thank the human participants for their time contribution. We acknowledge the support of the Australian Signals Directorate, the ASD-ANU Co-Lab, and Catherine Travis to the first author's thesis. We would also like to thank Associate Professor Alice Richardson from the ANU Statistical Consulting Unit for her statistical insights and expertise, which greatly assisted in this research.

6. References

[1] Deshmukh, N., Duncan, R. J., Ganapathiraju, A. and Picone, J. "Benchmarking human performance for continuous speech recognition", ICSLP'96 Fourth International Conference on Spoken Language Processing Proc., 1996.

[2] Goldwater, S., Jurafsky, D. and Manning, C. D. "Which words are hard to recognize? Prosodic, lexical, and disfluency factors that increase speech recognition error rates", Speech Communication, 52(3):181-200, 2010.

[3] Kong, X., Choi, J.-Y. and Shattuck-Hufnagel, S. "Evaluating automatic speech recognition systems in comparison with human perception results using distinctive feature measures", IEE International Conference on Acoustics, Speech and Signal Processing, 2017.

[4] Richter, C., Feldman, N. H., Salgado, H. and Jansen, A. "Evaluating low-level speech features against human perceptual data", Transactions of the Association for Computational Linguistics, 5:425-440, 2017.

[5] Benkí, J. R. "Place of articulation and first formant transition pattern both affect perception of voicing in English", Journal of Phonetics, 29(1):1-22, 2001.

[6] Dmitrieva, O., Llanos, F., Shultz, A. A. and Francis, A. L. "Phonological status, not voice onset time, determines the acoustic realization of onset f0 as a secondary voicing cue in Spanish and English", Journal of Phonetics, 49:77-95, 2015.

[7] Liberman, A. M., Delattre, P. C. and Cooper, F. S. "Some cues for the distinction between voiced and voiceless stops in initial position", Language and Speech, 1(3):153-167, 1958.

[8] Lisker, L. "Is it VOT or a first-formant transition detector?", The Journal of the Acoustical Society of America, 57(6):1547-1551, 1975.

[9] Repp, B. H. "Relative Amplitude of Aspiration Noise as a Voicing Cue for Syllable-Initial Stop Consonants", Language and Speech, 22(2):173-189, 1979.

[10] Bastian, J., Delattre, P. and Liberman, A. M. "Silent interval as a cue for the distinction between stops and semivowels in medial position", The Journal of the Acoustical Society of America, 31(11):1568-1568, 1959.

[11] Cooper, F. S., Delattre, P. C., Liberman, A. M., Borst, J. M. and Gerstman, L. J. "Some experiments on the perception of synthetic speech sounds", Journal of the Acoustical Society of America, 24(6):597-606, 1952.

[12] Dorman, M. F., Raphael, L. J. and Liberman, A. M. "Some experiments on the sound of silence in phonetic perception", Journal of the Acoustical Society of America, 65(6):1518-1532, 1979.

[13] Liberman, A. M., Delattre, P. and Cooper, F. S. "The role of selected stimulus-variables in the perception of the unvoiced stop consonants", American Journal of Psychology, 497-516, 1952.

[14] Repp, B. H. "Phonetic trading relations and context effects: New experimental evidence for a speech mode of perception", Psychological Bulletin, 92(1):81, 1982.

[15] Kronrod, Y., Coppess, E. and Feldman, N. H. "A unified account of categorical effects in phonetic perception", Psychonomic Bulletin & Review, 23(6):1681-1712, 2016.

[16] Boersma, P. "Praat, a system for doing phonetics by computer", Glot International, 5:9/10: 341-345, 2001.

[17] Qualtrics. [Software]. http://www.qualtrics.com. 2017.

[18] Mozilla. Project DeepSpeech. Retrieved from https://github.com/mozilla/DeepSpeech 2020.

[19] Google. Speech-to-Text. Retrieved from https://cloud.google.com/speech-to-text/ 2020.

[20] Repp, B. H. "Closure Duration and Release Burst Amplitude Cues to Stop Consonant Manner and Place of Articulation", Language and Speech, 27(3):245-254, 1984.

[21] Bastian, J., Eimas, P. D. and Liberman, A. M. "Identification and discrimination of a phonemic contrast induced by silent interval", Journal of the Acoustical Society of America, 33(6):842-842, 1961.

[22] Harris, K. S., Bastian, J. and Liberman, A. M. "Mimicry and the perception of a phonemic contrast induced by silent interval: Electromyographic and acoustic measures", Journal of the Acoustical Society of America, 33(6):842-842, 1961

[23] Belinkov, Y. and Glass, J. "Analyzing hidden representations in end-to-end automatic speech recognition systems", Advances in Neural Information Processing Systems Conference, 2017.

[24] Nagamine, T., Seltzer, M. L. and Mesgarani, N. "Exploring how deep neural networks form phonemic categories", 16th Annual Conference of the International Speech Communication Association, 2015.

[25] Pellegrini, T. and Mouysset, S. "Inferring phonemic classes from CNN activation maps using clustering techniques", 17th Annual Conference of the International Speech Communication Association, 2016.

[26] Morais, R. "A Journey to <10% Word Error Rate". Retrieved from https://hacks.mozilla.org/2017/11/a-journey-to-10-word-error-rate/ 2017.

[27] Fant, G. Acoustic Theory of Speech Production, Mouton, 1960.

[28] Liberman, A. M. and Mattingly, I. G. "The motor theory of speech perception revised", Cognition, 21(1):1-36, 1985.

[29] Toscano, J. C. and McMurray, B. "Cue Integration with Categories: Weighting Acoustic Cues in Speech Using Unsupervised Learning and Distributional Statistics", Cognitive Science, 34(3):434-464, 2010.

[30] Hermansky, H. "Coding and decoding of messages in human speech communication: Implications for machine recognition of speech", Speech Communication, 106:112-117, 2019.

[31] Whalen, D. H., Gick, B., Kumada, M. and Honda, K. "Cricothyroid activity in high and low vowels: exploring the automaticity of intrinsic F0", Journal of Phonetics, 27(2):125-142, 1999.

Rhoticity and hiatus breaking in Australian English: Associations with community diversity

Andy Gibson, Joshua Penney, Felicity Cox

Centre for Language Sciences, Department of Linguistics, Macquarie University

andy.gibson@mq.edu.au, joshua.penney@mq.edu.au, felicity.cox@mq.edu.au

Abstract

Australian English is considered non-rhotic, however increasing ethnolinguistic diversity may lead some speakers to exhibit partial rhoticity, particularly in areas of high diversity. r-sandhi is a phonological phenomenon that differentiates rhotic and non-rhotic varieties allowing realisations of /ɹ/ to arise in V_1#V_2 hiatus in non-rhotic speech whilst being absent in non-prevocalic contexts. V_1#V_2 hiatus can also be resolved by glottalisation/glottal stop insertion. We present an auditory analysis showing that children and teenagers from diverse areas are more likely to exhibit instances of non-prevocalic-r and are also more likely to use glottalisation to mark word boundaries in potential r-sandhi contexts.

Index Terms: rhoticity, vowel hiatus, linking-r, glottalisation, diversity, multicultural Australian English

1. Introduction

1.1. Multicultural Australian English

Australia is one of the most ethnically diverse countries in the world with Sydney its most multicultural city [1]. The 2016 census found that 42.9% of people in Greater Sydney were born overseas with 66.9% of people having at least one parent born overseas and 38.2% of households using a non-English language in the home [2]. These statistics do not reveal the true scope of diversity across the city which varies greatly from extremely diverse areas like Auburn where a language other than English is spoken in 83.5% of households compared with only 12.2% in Pittwater.

Across two ongoing projects we are exploring the role that ethnolinguistic diversity plays in the speech of Australian English (AusE) speaking young people. Ethnolinguistic diversity has been underexamined in AusE with the vast majority of studies focusing on the mainstream variety. These projects are designed to help fill this gap in our knowledge of variation in present-day AusE and provide a more representative picture of the variety. The Child Speech, Community Diversity (CSCD) project is longitudinally tracking the speech of children from both Sydney and New South Wales more broadly, from preschool through to their second year of school. The Multicultural Australian English (MAE) project studies the speech of teenagers from a range of areas across Sydney.

Here we focus on a feature of AusE that is typically described as one of its defining characteristics: non-rhoticity [3–5]. We examine the incidence of rhoticity according to community diversity in two controlled datasets (children and teenagers) and then explore how community diversity factors into the management of word-boundary phonological strategies.

1.2. Rhoticity and linking-r

Rhotic and non-rhotic varieties of English are differentiated by the presence of non-prevocalic-r in the former and its absence in the latter. Southern Standard British English (SSBE) and Southern Hemisphere varieties of English (such as New Zealand English (NZE) and AusE) are typically described as non-rhotic (but see [6] for NZE), although anecdotal reports suggest rhoticity may be present in the speech of some AusE speakers from diverse communities. An associated phenomenon that differentiates rhotic and non-rhotic dialects is r-sandhi, the insertion of a realisation of /ɹ/ at sites of potential vowel hiatus in non-rhotic varieties. A phonologically conditioned liaison strategy is used whereby /ɹ/ may arise in V_1#V_2 hiatus contexts (e.g., *four eggs, four o'clock*) when V_1 is non high, though /ɹ/ is absent in non-prevocalic environments (e.g., *four dogs*).

In a study of New Zealand English, [7] found that the use of linking-r at word-internal morpheme boundaries is near-categorical. At word boundaries, however, the potential V_1#V_2 hiatus can be resolved with either linking-r or glottalisation/glottal stop. Several studies have shown that the choice between linking-r and glottalisation is prosodically conditioned, primarily by the prominence of V_2 [3–5]. The choice of hiatus breaking strategy often depends on the strength of the prosodic boundary [3–5, 8–10] with glottalisation and linking-r appearing in complementary distribution. When V_2 is stressed, glottalisation is more likely and when V_2 is unstressed linking–r is more likely. [11] proposed the glottal stop to be the optimal boundary marking segment in V_1#V_2 contexts as it is maximally consonantal compared to an approximant which minimises contrast between surrounding vowels. This echoes findings from other studies (e.g., [12–14]) that show increased use of glottalisation when a stressed vowel appears to the right of a boundary.

The use of glottalisation rather than linking-r may also be sociolinguistically conditioned. Studies have found that speakers from some diverse communities in Britain are more likely to use glottalisation than linking-r in hiatus contexts. [15] and [16] showed that amongst speakers from Tower Hamlets in London, boys with a Bangladeshi background used a greater percentage of glottalisation in linking-r contexts than Anglo-background boys and girls from the same area, who had high rates of linking-r. Other studies of hiatus resolution (not specific to the linking-r context, e.g. [17–19]) have also found that speakers from high diversity areas make greater use of glottalisation to resolve hiatus. In a corpus of pop and hip hop songs, [20] found that linking-r occurred at lower rates amongst African American performers, who tended to use glottalisation, while American performers of European descent used high

rates of linking-r. As in other studies, a stressed V_2 favoured the use of glottalisation while an unstressed V_2 favoured linking-r across all social groups.

1.3. Research questions

If some AusE speakers in diverse communities are partially rhotic, as anecdotal evidence suggests, the question arises as to whether and how word boundary marking strategies might also vary across communities. We explore this question through an auditory analysis of speech data collected from children and teenagers, from areas that vary widely in their levels of ethnolinguistic diversity, elicited through a picture naming task designed to sample various phonological contexts including potential environments for both non-prevocalic-r and linking-r.

Our research aims to systematically assess whether rhoticity is present in the speech of the Australia-born children and teenagers in our two studies, and if so, whether it occurs more in areas with greater diversity, and in particular vocalic environments.

We then look at the way potential vowel hiatus at word boundaries is managed by the speakers in this dataset. We expect to replicate the prosodic effect described above: greater use of glottalisation/glottal stops at strong boundaries (i.e., when the vowel on the right edge of the potential hiatus is stressed) and greater use of linking-r at weaker boundaries (i.e., when the vowel on the right of the potential hiatus is unstressed). We will also explore whether the management of potential vowel hiatus at word boundaries varies with respect to community diversity as has been suggested in [15–19].

2. Analysis of rhoticity

2.1. Rhoticity Methods

For both the CSCD and MAE datasets, an impressionistic analysis of all potential non-prevocalic-r tokens was conducted for the presence/absence of rhoticity (e.g., in words like *car, shirt, water)*. All non-final schwa environments and function words were removed. The analysis was conducted in a quiet environment utilising headphones and Praat software [21], with reference made to spectrograms and waveforms where necessary.

2.1.1. CSCD methods

Here we include only data from the first timepoint of our longitudinal project, when most children were in their last year of preschool. The median age was 5;0 (ranging between 4;9 and 6;3, with a total of 57 females and 78 males). All children included in this analysis were born in Australia and live in New South Wales, with a large proportion from Sydney, and a cluster from the Mid North Coast. For this analysis a total of 135 children from a range of areas varying according to their level of ethnolinguistic diversity were included.

Children engaged in a self-recorded picture naming task delivered via the Gorilla online platform [22], framed as a game where the child helped a cartoon alien find its friends and its spaceship. 150 single words and short phrases were elicited but incidental items were often recorded as the children engaged with the task. Self-recorded in their own homes, the recordings come from a wide range of devices of varying quality, with a sample rate of either 44.1kHz or 48kHz.

The 1st author and a research assistant coded the majority of the tokens, with the remainder coded by the 3rd author. In order to check for reliability, 7% of tokens were coded by more than one analyst. While the results from later timepoints are not presented in this paper, our inter-coder reliability rate from the full dataset (three timepoints) was 96%.

2.1.2. MAE methods

The data analysed here were collected via a picture naming task in which 225 single words and short phrases were elicited through presentation of images on a computer monitor. Speakers were recruited from schools in disparate areas of Sydney that differ both in terms of the level of diversity as well as the dominant non-English languages spoken. Data from 117 adolescent speakers (female: n = 65; male: n = 52) aged between 15–17 years are included here. 69 of the speakers were recorded face-to-face in a quiet room in their school using a Zoom H6 recorder with a sample rate of 44.1kHz; 48 speakers were recorded remotely via a supervised video call while at school using a browser-based recorder [23] with a sample rate of 48kHz. These files were subsequently resampled to 44.1kHz.

The third author coded the majority of tokens, with the remainder coded by the second author. 10% of the items were coded by both analysts, with an inter-coder agreement rate of 96%.

2.2. Rhoticity Results

8643 tokens (children: n = 4548; teenagers: n = 4095) of potential non-prevocalic-r were analysed. A total of 694 tokens (8%) of potential occurrences were deemed to be rhotic (7.2% rhoticity in children's data and 8.9% for teenagers). 39% (53/135) of children and 33% (39/117) of teenagers produced at least one token of non-prevocalic-r. 19% of the children (26/135) and 19% of the teenagers (22/117) had a rate of 10% rhoticity or more. In each of the two groups three speakers had a rate of rhoticity of over 80%.

To examine the role of community diversity, we use a postcode-based measure, obtained from census data using Table Builder [2]. Modeling the variation in our results according to the percentage of households in a given postcode area that speak a language other than English, we treat community diversity as a continuous variable. Note that gender differences will not be examined in this paper.

We fit a binomial generalised linear mixed effects regression model predicting the log-odds of non-prevocalic-r, with speaker and word as random intercepts, and three fixed effects:

- diversity (the proportion of households in the speaker's postcode area that speak a language other than English, centred and scaled);
- cohort (children vs. teenagers);
- vocalic context (a six-level factor distinguishing the lexical sets NEAR, SQUARE, NURSE, START, NORTH, and LETTER, i.e., word-final instances of schwa with orthographic <r>, e.g., *water*).

An interaction between diversity and cohort was tested and was not significant. Results show that speakers from more diverse areas were more likely to produce non-prevocalic-r than speakers from more homogeneous areas (Estimate=2.067; $p<0.001$). The vowel context was highly significant ($p<0.001$), with three r-favouring contexts (NEAR, SQUARE and NURSE) all being significantly more likely to exhibit rhoticity than three r-disfavouring contexts (START, NORTH and LETTER). Finally, there was a significant effect of cohort, such that the children were more likely to produce non-prevocalic-r than the teenagers (Estimate=- 1.510; $p=0.004$). The effect of community diversity is plotted in Figure 1, which shows the log-odds of rhoticity,

back-transformed to percentages according to diversity (also back-transformed) along with points for each speaker's rate of rhoticity, coloured according to cohort.

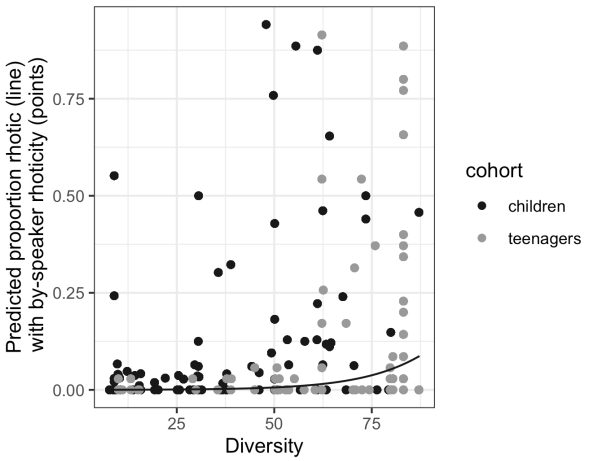


Figure 1: *Predicted rhoticity proportions according to community diversity (line), with raw by-speaker rhoticity proportions for children (black points) and teenagers (grey points).*

2.3. Rhoticity Discussion

The results provide evidence of variable rhoticity in Australia-born speakers. There is a clear effect of community diversity with higher rates of rhoticity in more diverse areas, defined as those areas where there are fewer households where only English is spoken, with children being more likely to produce rhoticity than teenagers, once the community diversity effect is held constant.

The likelihood of non-prevocalic-r also varies according to the vocalic context, with greater rhoticity in NEAR, SQUARE and NURSE than other environments.

It is crucial to note here that these data come from a picture naming task. Items are therefore in citation form in a performative/game-like context which may promote the use of clear speech or perhaps exaggerated productions in some speakers. Future work will assess rhoticity rates in the spontaneous speech data collected for the teenagers.

We have defined community diversity according to the proportion of households in a given area that speak languages other than English. While all participants were born in Australia, English is not the dominant language used in the home for many of these speakers. We acknowledge the effect of individuals' language backgrounds in addition to these area-based measures. However, since language background and our diversity measure are highly correlated, it is not statistically feasible to include both measures within the same analysis. Ultimately, it is important to disentangle the effects of an individual's language background and the ethnolinguistic diversity of the area in which they live. Future analyses will consider the role of language background and whether rhoticity can be attributed to transfer from a heritage language or dialect. The finding that children are more likely than teenagers to produce rhoticity may be due to greater influence of language used in the home on children's speech prior to starting school, after which peer influence gains momentum.

3. Analysis of glottalisation and linking-r at sites of potential vowel hiatus

Having found evidence for some rhoticity in Australia-born teenagers and children, particularly in diverse areas, we are drawn to the question of hiatus management. While both rhotic and non-rhotic speakers may produce /ɹ/ at word boundaries in phrases such as *four apples*, they may do so for different reasons, with the /ɹ/ having special status as part of a hiatus management system for non-rhotic speakers. We now turn to an analysis of the strategies employed by speakers at sites of potential vowel hiatus, exploring first the full range of realisations that occur, and then turning to a systematic analysis of the two primary variants: linking-r and glottalisation.

3.1. Hiatus methods

Both CSCD and MAE employed a counting task in the picture naming game where the speakers were encouraged to produce two-word phrases. For the children, this included four phrases that contain $V_1\#V_2$ hiatus (*four* + *apple/egg/alien/o'clock*), three with a stressed V_2 and one with an unstressed V_2 (*o'clock*). Similar two-word phrases were also elicited from the teenagers, as well as a task designed to elicit the same nouns paired with the possessive determiner *her*. Ten separate contexts were elicited: six with a stressed vowel on the right edge of the hiatus (*ear, eye, arm, eagle, oar, uber*) and 4 with a weak V_2 (*alarm, exam, award, o'clock*). Both sets of data were coded for the presence/absence of linking-r and the presence/absence of glottalisation/glottal stop. Glottalisation and full glottal stops are treated as equivalent for this analysis [24]. An important difference in methods between the cohorts is that for the children's data we included cases of labiovelar or labiodental glides within the linking-r category. In these tokens, the child appears to be enacting 'glide insertion' for the purposes of breaking the hiatus despite not being able to produce an adult-like /ɹ/.

All tokens were coded by at least two analysts, with discrepancies checked and resolved by a third analyst.

3.2. Hiatus results

For the children, a total of 359 tokens were analysed, revealing three main variants: true vowel hiatus (where there was no evidence of either linking-r or glottalisation) occurred in 29.5% of tokens; linking-r occurred in 34%, and glottalisation in 35.4% of tokens.

For the teenagers, a total of 2050 tokens were analysed (*four*: 1142; *her*: 908). As no differences were found according to the strength of the left-hand environment (strong *four* vs. weak *her*), these contexts are combined to focus on differences according to the strength of the right-hand environment. The two primary variants were linking-r and glottalisation. In addition, there were a number of other realisations, including a combination of rhoticity and glottalisation, true vowel hiatus, glide insertion ([w] or [j]), and elision of the right edge vowel (which only occurred in the tokens including *o'clock*). These realisations are excluded from the statistical analysis presented below, which focuses on the binary distinction between glottalisation and linking-r.

For teenagers from ethnolinguistically homogeneous areas (those where less than 20% of households speak a language other than English) there is noticeable systematicity to the resolution of potential vowel hiatus at word boundaries. When the right-edge vowel is stressed, glottalisation is used in 88% of

tokens. When the right-edge vowel is unstressed, linking-r is the preferred variant (72%).

This is not the case in more diverse areas (postcodes where more than 20% of households speak a language other than English). At stressed boundaries, glottalisation is near-categorical (95%). At weak boundaries, however, the use of both glottalisation (46%) and linking-r (40%) occur at similar rates, and this is the only environment where true vowel hiatus occurs in a sizeable number of tokens (13%). It was also in this environment where most of the variants excluded from the analysis (such as elision) occurred.

Figure 2 summarises these raw results, showing the patterning of the three main variants (glottalisation, linking-r and true vowel hiatus) for each cohort (children and teenagers) according to the stress of the right edge vowel and whether the speaker lives in a more homogeneous area or a more diverse area.

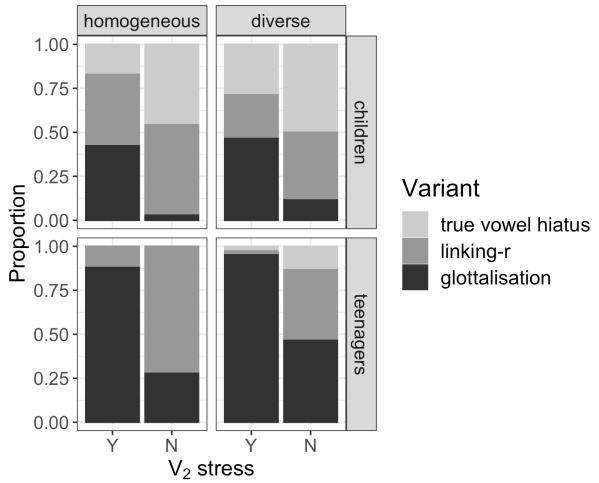


Figure 2: *Raw proportions of the three main variants at sites of potential vowel hiatus according to cohort, community diversity and the stress of the right-edge vowel.*

For the purposes of modeling the data as a binary variable, all tokens of true vowel hiatus were removed from the dataset, along with the small number of other infrequent variants. A binomial linear mixed effects model was fit to the dataset (1810 tokens across 194 speakers), predicting the log-odds of glottalisation (vs. linking-r), with speaker as a random intercept and three fixed effects:

- diversity (the proportion of households in the speaker's postcode area that speak a language other than English, centred and scaled);
- cohort (children vs. teenagers);
- stress of the vowel on the right-edge of the potential hiatus.

The three main effects were significant. Speakers from more diverse areas were more likely to produce glottalisation than speakers from more homogeneous areas (Estimate=1.135; $p<0.001$). Glottalisation was significantly less likely when V_2 was unstressed (Estimate=-4.525; $p<0.001$), and teenagers were more likely to use glottalisation than children (Estimate=3.086; $p<0.001$). The interaction between diversity and stress described above with respect to the raw data was tested but was not significant.

3.3. Hiatus discussion

An analysis of potential environments for linking-r revealed different strategies for managing hiatus between the children and the teenagers. While true vowel hiatus was rare for the teenagers, it was common for children to use neither linking-r nor glottalisation at the word boundary. Linking-r and glottalisation were the primary variants for the teenagers, and across the whole dataset the choice between these variants was strongly motivated by the prosodic context consistent with previous studies [e.g., 3, 4, 5]. When the vowel on the right edge of the potential hiatus is stressed, glottalisation is preferred. Linking-r mainly occurs when V_2 is unstressed.

Additionally, we find again a significant effect of community diversity, with those in more diverse areas more likely to use glottalisation to resolve vowel hiatus than those in more homogeneous areas. While the interaction of stress with community diversity was not significant, teenagers from homogeneous areas seemed to be particularly systematic in their use of glottalisation at strong boundaries and linking-r at weak boundaries.

Finally, teenagers used more glottalisation than children. Given the high rate of true vowel hiatus for children, this difference may due to developmental factors. The three variants observed here can be viewed as falling on a scale from true vowel hiatus ($V_1\#V_2$) – the absence of a boundary, to linking-r – a medium-strength boundary, to glottalisation – a strongly marked boundary. The children may not yet have learned the norms for breaking vowel hiatus at word boundaries. Ongoing work will assess whether these children develop in the direction of the teenagers in their first years of schooling.

4. General discussion

This analysis has assessed the role of community diversity on the realisation of two distinct but related phenomena. We found evidence that partial rhoticity is used by Australia-born children and teenagers, particularly in areas with greater ethnolinguistic diversity. At sites of potential vowel hiatus, linking-r is traditionally thought to be the primary hiatus-breaking device in non-rhotic dialects [3]. Consistent with [3–5], we find that for $V_1\#V_2$ contexts, this is much more likely when V_2 is a weak syllable, and glottalisation is strongly favoured when V_2 is stressed.

We might expect that the higher incidence of rhotic speakers in more diverse areas would lead to greater use of linking-r, however the present results suggest that speakers in more diverse areas use glottalisation to mark word boundaries to a greater extent than those in homogeneous areas, where linking-r and glottalisation appear to be in complementary distribution, governed by the strength of the boundary. It may be the case that strengthening the word boundary may offer communicative enhancement in areas where speech patterns in the ambient language of the community are extremely variable. This suggestion requires empirical support.

Future analyses will examine the acoustic characteristics of the rhotic consonants to provide a more nuanced account of variation in this highly diverse sample of speakers.

Much sociophonetic work describing AusE has been focused on monolingual speakers of Anglo-Celtic heritage from relatively homogeneous communities. By systematically studying speakers from areas with greater ethnolinguistic diversity, we are able to trace the development of variation and change in AusE that reflects the increasing diversity of its speakers.

5. References

[1] Australian Bureau of Statistics: 'Cultural Diversity: Census of Population and Housing: Australia Revealed, 2016.', 2017

[2] Australian Bureau of Statistics: 'Census 2016', 2016

[3] Cox, F., Palethorpe, S., Buckley, L., and Bentink, S.: 'Hiatus resolution and linking 'r' in Australian English', Journal of the International Phonetic Association, 2014, 44, (2), pp. 155–178

[4] Yuen, I., Cox, F., and Demuth, K.: 'Planning of Hiatus-Breaking Inserted /ɹ/ in the Speech of Australian English–Speaking Children', Journal of Speech, Language, and Hearing Research, 2017, 60, (4), pp. 826–835

[5] Yuen, I., Cox, F., and Demuth, K.: 'Prosodic effects on the planning of inserted /ɹ/ in Australian English', Journal of Phonetics, 2018, 69, pp. 29–42

[6] Marsden, S.: 'Are New Zealanders "rhotic"?', English World-Wide, 2017, 38, (3), pp. 275–304

[7] Hay, J., Drager, K., and Gibson, A.: 'Hearing r-sandhi: The role of past experience', Language, 2018, 94, (2), pp. 360–404

[8] Foulkes, P.: 'Rule inversion in a British English dialect - a sociolinguistic investigation of [r]-sandhi in Newcastle upon Tyne', University of Pennsylvania Working Papers in Linguistics: A Selection of Papers from NWAV 25, 1997, 4, (1), pp. 259–270

[9] Mompeán, J. A., and Gómez, F. A.: 'Hiatus Resolution Strategies in Non-rhotic English: The Case of /r/-Liaison', in Lee, W.S. and Zee, E. (Eds.), Proceedings of 17th International Congress of Phonetic Sciences, Hong Kong 2011, pp. 1414–1417

[10] Mompeán, J. A.: '/r/-sandhi in the speech of Queen Elizabeth II', Journal of the International Phonetic Association, 2021, pp. 1–32

[11] Uffmann, C.: 'Intrusive [r] and optimal epenthetic consonants', Language Sciences, 2007, 29, (2), pp. 451–476

[12] Dilley, L., Shattuck-Hufnagel, S., and Ostendorf, M.: 'Glottalization of word-initial vowels as a function of prosodic structure', Journal of Phonetics, 1996, 24, (4), pp. 423–444

[13] Garellek, M.: 'Glottal stops before word-initial vowels in American English: distribution and acoustic characteristics', UCLA Working Papers in Phonetics, 2012, 110, (1), pp. 1–23

[14] Garellek, M.: 'Voice quality strengthening and glottalization', Journal of Phonetics, 2014, 45, pp. 106–113

[15] Britain, D., and Fox, S.: 'The Regularisation of the Hiatus Resolution System in British English. A Contact-Induced 'Vernacular Universal'?', in Filppula, M., Klemola, J., and Paulasto, H. (Eds.): 'Vernacular Universals and Language Contacts: Evidence from Varieties of English and Beyond.' (Routledge, 2008), pp. 177–205.

[16] Fox, S.: 'The New Cockney: New Ethnicities and Adolescent Speech in the Traditional East End of London' (Palgrave Macmillan UK, 2015)

[17] Cheshire, J., Kerswill, P., Fox, S., and Torgersen, E.: 'Contact, the feature pool and the speech community: The emergence of Multicultural London English', Journal of Sociolinguistics, 2011, 15, (2), pp. 151–196

[18] Meyerhoff, M., Birchfield, A., Ballard, E., Charters, H., and Watson, C.: 'Definite change taking place: Determiner realization in multiethnic communities in New Zealand', University of Pennsylvania Working Papers in Linguistics, 2020, 25, (2), pp. 71–78

[19] 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, 2022, pp. 1–31

[20] Gibson, A.: 'Sociophonetics of Popular Music: Insights from Corpus Analysis and Speech Perception Experiments', PhD dissertation, University of Canterbury, 2019. [Online]. Available: https://ir.canterbury.ac.nz/handle/10092/17892

[21] Boersma, P. and Weenink, D.: 'Praat: doing phonetics by computer', Version 6.2.19, retrieved 21 September 2022 from http://www.praat.org

[22] Anwyl-Irvine, A.L., Massonnié, J., Flitton, A., Kirkham, N., and Evershed, J. K.: 'Gorilla in our midst: An online behavioral experiment builder', Behavior Research Methods, 2020, 52, (1), pp. 388–407

[23] German Research Centre for Artificial Intelligence (DFKI): 'speech-to-flac', retrieved 21 September 2022 from https://github.com/mmig/speech-to-flac

[24] Penney, J., Cox, F., and Szakay, A.: 'Effects of glottalisation, preceding vowel duration, and coda closure duration on the perception of coda stop voicing', Phonetica, 2021, 78, (1), pp. 29–63

Longevity of an ethnolectal marker in Australian English: word-final (er) and the Greek-Australian community

Elena Sheard

ARC Centre of Excellence for the Dynamics of Language, Australian National University

elena.sheard@anu.edu.au

Abstract

This paper builds on a prior study of (er) lengthening in Australian English that demonstrated (er) in prosodically final position (*other, remember*) has displayed incremental lengthening between the 1970s and 2010s, a change led by Greek Teenagers in the 1970s. Here, I extend the analysis to subsequent generations of Greek Australians. Analyses of 3941 prosodically final (er) tokens reveal that Greek-Australian Adults born in the 1960s maintain their leading role in the lengthening of (er). Young Adults born in the 1990s, however, display diminished ethnic differentiation, demonstrating limited longevity of the ethnolectal status of (er).

Index Terms: ethnic variation, Australian English, sociophonetics, word-final (er), Greek Australians

1. Introduction

Ethnic minorities have long been documented as participants in language variation. They have also been identified as potential leaders of language change in multilingual urban centres such as London [1], Gothenburg [2], and Sydney [3]. This study shifts focus to the longevity of ethnic differentiation for a known ethnolectal feature of Australian English: word-final (er). While we know that Greek-Australians led the lengthening of (er) in the 1970s [4], we do not know whether they continue to be differentiated from other ethnic communities in the 2010s. This analysis tracks the Greek-Australian community over two time points, three age groups, and two generations. When combined with comparable data from Italian- and Anglo-Australians, we can further understand the role of ethnic minorities in language change, and more accurately assess the ethnolectal status of (er) over time.

In 2020, Grama, Travis and Gonzalez [4] conducted a real-time analysis of 193 socially stratified Australians of Anglo, Greek, Italian, and Chinese background as part of the Sydney Speaks project. This analysis covered three generations and four age groups of speakers (Adults born in the 1930s, Teens and Adults born in the 1960s and Young Adults born in the 1990s) and identified that (er) lengthening progressed incrementally over time. This change was restricted to Intonation-Unit (IU) final position, and was conditioned by ethnicity, class, and gender. Specifically, the variable first lengthened in the speech of Greek- and Italian-Australian Teenagers in the 1970s, followed by Working Class Anglo- and Italian-Australian Adult women, before being taken up by the rest of the community in the 2010s. The Anglo-Australian community effectively caught up to the benchmark set by the Greek- and Italian-Australian 1970s Teens.

These findings were consistent with previous research that has identified (er) as an ethnically marked variable. For example, Clyne, et al. [5] identified an open final vowel in /iə/ as one of the seven 'most conspicuous phonological features' of the Greek ethnolect. Warren [6] also noted the widespread realisation of (er) as '[a] in final syllables' as a salient feature of 'ethnic' Australian English. Acoustic analyses of (er) realisation in Australian English have also identified ethnic differentiation [e.g. 7, 8]. Kiesling's [7] analysis of 6 Anglo and 15 non-Anglo speakers (including 7 Greek- and 2 Italian-Australians) indicated that the 'significant pronunciation difference between Anglo and non-Anglo speakers', was most pronounced for Greek-Australians, who produced a backer and longer vowel, often in combination with High Rising Tone. As noted in Grama, Travis and Gonzalez [4], representations of (er) in Australian media have also supported an association with the Greek-Australian community.

Grama, Travis and Gonzalez [4] hypothesised, based on the patterning of the three age groups, that the ethnic differentiation in (er) realisation had diminished in the 2010s. However, data from the Greek-Australian community in the 2010s was not available at the time of writing. This leaves a gap in our knowledge about the status of (er) as an ethnolectal marker of the Greek-Australian community over time, and the role of this community in progressing this change in the 2010s. This study addresses this gap by extending the analysis of previously existing Sydney Speaks data to new data from Greek Australian Adults (n = 13) and Young Adults (n = 6). By including this new data, we can track the Greek-Australian community relative to their Anglo- and Italian-Australian peers in real and apparent time across two time points (1970s and 2010s), three age groups (1970s Teens, 2010s Adults and Young Adults) and two generations (1970s Teens and 2010s Adults born in the 1960s, 2010s Young Adults born in the 1990s). We can therefore identify whether Greek-Australians in Sydney remain ahead in this change and, by extension, assess the longevity of (er) as an ethnolectal marker of this community.

2. Methods

Word-final (er) is here defined as 'unstressed /ə/ in word-final position in minimally disyllabic words with a following *r* in the orthography', as in [4]. The lengthening of (er) has occurred specifically in open syllables in prosodically final position, with this degree of lengthening not shared by tokens in prosodically medial position [4]. This variable will henceforth be referred to as (er), noting that it occurs specifically in word-final position.

2.1. Corpus and participants

Participants are drawn from a corpus of sociolinguistic interviews conducted with 171 younger and older speakers over two time periods, as part of the Sydney Speaks project [9]. The first group of interviews were recorded in the late 1970s for the Sydney Social Dialect Survey (SSDS), conducted by Barbara

Horvath [3]. Here, I draw on those interviews with Teenagers born in the 1960s. The second group of interviews are original recordings made by the Sydney Speaks Project in the 2010s with Adults and Young Adults (born 1960s and 1990s, respectively). The composition of the corpus reported on here is summarised in Table 1.

Participants are Sydneysiders of Anglo-Celtic, Italian, and Greek background. The Anglo-Australian participants' parents and grandparents are known to have grown up in Australia. The non-Anglo-Australian participants were born in Australia, or arrived before the age of five, and their parents migrated from Italy or Greece. The corpus is stratified according to age, gender, and socio-economic status. Speakers were assigned to three socio-economic status groups based on Hierarchical k-means cluster analyses of measures for level of occupation, education, suburb, and high school type (see [4] for details). Following [3], these groups are referred to as Lower Working, Upper Working and Middle Class (LWC, UWC, MC).

These two corpora allow us to observe ethnicity-based language variation and change in real time. This is highly valuable given our currently limited understanding of the longevity of ethnolectal features across generations.

Table 1 *Demographic breakdown of participants*

	1970s		**2010s**				
	Teens		*Adults*		*Young Adults*		
	M	F	M	F	M	F	
Anglo	12	12	10	10	12	10	66
Greek	10	13	5	8	4	2	42
Italian	13	12	11	13	8	6	63
Total	35	37	26	31	24	18	171

2.2. Data preparation, token extraction and analysis

The extraction of (er) tokens was based on two levels of transcription in ELAN [10]. In the first level, orthographic transcriptions of the sociolinguistic interviews were force-aligned in LaBB-CAT [11]. In the second level, prosodic information was included. Transcriptions were broken down into Intonation Units (IUs), here defined as 'a stretch of speech uttered under a single, coherent intonation contour', following the protocols of Du Bois, et al. [12].

Following the protocol in [4], each instance of (er) from the force-aligned transcriptions was matched with the corresponding prosodic transcription to determine the token's position in the IU (final or medial) and intonation contour ('transitional continuity' as defined in [12]). An example of IU-final (er) is presented in (1)

(1) AMAR: *so we all went to church that night* ***together***.
SydS_GOM_144_Amar, 1853.89

As outlined in [4], the analysis focused on (er) in open syllables regardless of intonation contour or morphological status. This is because the effect of intonation is consistent across age groups (i.e., the change towards longer (er) over time was not driven by a single intonation contour), and the realisation of (er) duration is not affected by whether –er is a separate morpheme or not (e.g., *remember* vs *teacher*). But, tokens with additional morphology (e.g., plural /s/) were excluded as they are not subject to lengthening. Each included token of (er) was coded according to speaker, speech rate, gender, community, age, and IU position.

Duration of (er) was measured based on the forced alignments produced in LaBB-CAT. Due to the relative accuracy of LaBB-CAT's boundary placement [13], as confirmed for (er) by spot-checking [4], tokens were not manually checked. Following [4], speech rate was calculated based on the number of syllables per minute, as calculated in LaBB-CAT, without further modification. As Grama, Travis, and Gonzalez [4, see Figure 5] demonstrated that a longer vowel was significantly correlated with a lower and backer realisation in the vowel space, the focus of this analysis is on (er) duration rather than F1 and F2 values.

A total of 3941 IU-final tokens were extracted for analysis. The data were examined visually, and observed patterns were further examined with three linear mixed-effects regression models. In the models fit to IU-final (er) tokens, one model included the 1970s Teens and 2010s Adults, focusing on the shift within this generation of speakers. The second model included 2010s Adults and Young Adults, focusing on the shift across generations. The third and final model analysed just the 2010s Young Adults to focus on ethnic differentiation within this age group. All models included speakers from all three ethnic communities.

Models were calculated using the *lmer* function as part of the *lme4* package [14] in *R* [15]. P-values were calculated using the *lmerTest* package [16]. For all models, duration of IU-final (er) was the dependent variable, speaker and word were included as random effects, and community, gender and speech rate were included as fixed effects. The first and second models included age as a fixed effect as they are focused on changes between age groups. The first model also included class as a fixed effect, but the uneven distribution of 2010s Greek participants made this inadvisable for the second and third models. For the same reason, the third model included just UWC and MC Young Adults. Final models were reached through pruning, with fixed effects tested and removed if not significant. Interactions between fixed affects were also tested for. Model fit was assessed based on the Akaike Information Criterion (AIC) through the *AICcmodavg* package [17] and where model fit was not significantly different, the simpler model was chosen.

3. Results

3.1. 1970s Teens to 2010s Adults

This section provides an overview of the variation and change for the generation of 1970s Teens and 2010s Adults, with speakers from both age groups all born in the 1960s. The final model for this generation returned the following significant fixed effects: community, gender, age, and speech rate. Consistent with the results reported by Grama, Travis and Gonzalez [4], the Greek-Australian 1970s Teens' (er) durations are significantly longer than both their Anglo and Italian peers, and the Italian-Australians are significantly longer than the Anglo-Australians.

Figure 1 depicts (er) duration (on the vertical axis) broken down by age (on the horizontal axis) with community represented by colour. For visualisation purposes, outliers above 400 ms in duration have been excluded from all Figures. As demonstrated in [4], the Anglo- and Italian-Australians significantly lengthen their realisations of (er), with the average duration respectively increasing 53% (73 to 112 ms) and 39% (91 to 126 ms) between the 1970s Teens and 2010s Adults.

Adding in the Greek-Australians demonstrates that they have undergone a similar degree of change in the same period. Their average (er) duration has increased 36% between the two age groups (112 to 152 ms, $df = 1350$, $t = 5.5$, $p < 0.0001$).

While the average (er) durations of all three groups have increased to a similar extent in milliseconds (respectively increasing by 39, 35 and 40 ms), the Anglo-Australians have proportionally increased the most.

But crucially, adding in the new data from the Greek-Australian community in the 2010s allows us to see that they have not only continued to progress the change, but they have also maintained their longer realisations of (er) relative to the other ethnic communities. Their realisation of (er) duration remains significantly different from their Anglo- ($df = 1440$, $t = -5.4$, $p < 0.0001$) and Italian-Australian ($df = 1440$, $t = -3.8$, $p < 0.001$) peers in the 2010s. For this generation, therefore, it would appear that (er) has maintained its status as an ethnolectal marker of the Greek-Australian community.

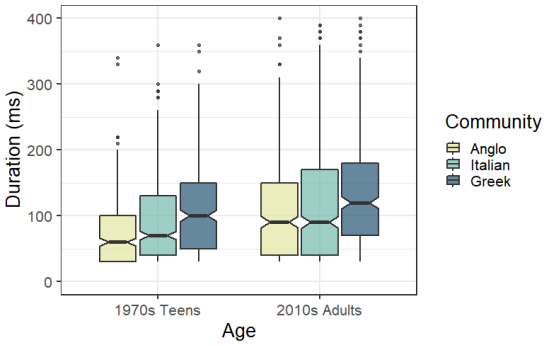


Figure 1 *(er) duration by age and ethnic community for 1970s Teens and 2010s Adults*

The changing class stratification of (er) for Greek-Australians is highly relevant to its ongoing status as an ethnolectal marker. Figure 2 similarly depicts (er) duration on the vertical axis and age on the horizontal axis. However, it presents the data specifically for the Greek-Australian 1970s Teens and 2010s Adults by class (LWC and UWC in orange, MC in purple). Consistent with [4], the LWC and UWC Greek-Australian 1970s Teenagers were most progressed in the lengthening of (er) (followed by Italian (male) Teenagers). It is clear from this figure that the Greek-Australian MC Teenagers were lagging in this change. In fact, they more closely resembled MC Anglo- and Italian-Australian Teenagers than the other Greek-Australians. In the 2010s, it is the MC Greek-Australians who have changed the most as Adults. The data therefore indicates that this class group has lengthened over time to meet the benchmark originally set by the rest of the Greek-Australian community in the 1970s. This has substantially contributed to the community's ongoing participation in this change and its differentiation from Anglo- and Italian-Australians in the 2010s.

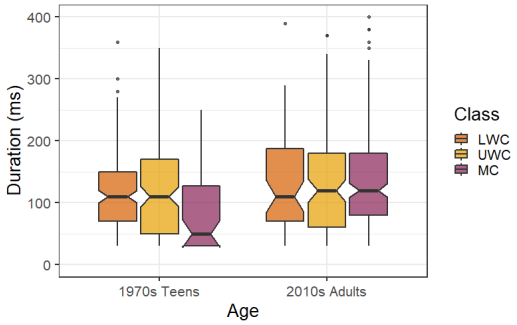


Figure 2 *(er) duration by age and class for Greek-Australian 1970s Teens and 2010s Adults*

Finally, the difference between men and women has increased over time for all three ethnic communities, with (Working Class) women leading the change (significant interaction between gender and age $df = 126$, $t = -2$, $p < 0.05$). Taking into account the comparatively greater lengthening of the Anglo LWC and the Italian WC 2010s Adult women [4], the ethnic differences in this generation have therefore contracted for women while being maintained for men.

The data indicates that the ongoing differentiation of the Greek-Australian Adults from their peers in the 2010s is not driven by a change or underlying difference in speech rate. There is a consistent effect across communities of faster speech rates significantly favouring shorter (er) duration. While this effect is more pronounced for the Greek-Australian 2010s Adults (significant interaction between speech rate, age, and community, $df = 2592.7$, $t = -2$, $p < 0.05$), there is no significant differences in speech rate between the three communities (based on a separate linear mixed effects model with speech rate as the dependent variable and an interaction between age and community as fixed effects). The independence of the Greek-Australian community's differentiation from speech rate is also supported by the fact that speech rate does not interact with community for the 1970s Teens, where the ethnic difference in duration is equally pronounced.

This data therefore captures a generation for which a linguistic variable is both an ethnolectal feature characteristic of a certain ethnic community, and a change in the process of being taken up by other ethnic groups in the wider speech community.

3.2. Introducing the 2010s Young Adults

This section introduces the 2010 Young Adults into the analysis to assess the longevity of (er) as an ethnolectal marker of the Greek-Australian community across generations. The Anglo-Australians have significantly lengthened their (er) duration between the two generations, while the Greek- and Italian-Australian Young Adults have not. This is reflected in a significant interaction between age and community for the Italian- ($df = 94$, $t = -2.6$, $p < 0.05$) and Greek- Australians ($df = 83$, $t = -2.1$, $p < 0.05$) returned by the final model for 2010s Adults and Young Adults. As a result, Young Adults from all three communities have similar median durations and the differentiation of the Greek-Australian community has diminished. This is shown in Figure 3, which is the same as Figure 1, but with the Young Adult Greek-Australians included and data presented separately for female and male speakers.

We can see from Figure 3 that the change in (er) duration between the 2010s Adults and 2010s Young Adults is relatively small compared to the change between the 1970s Teens and 2010s Adults. Most importantly, Greek-Australian Young Adults are not continuing to lengthen (er) further than the previous generation but are patterning with the other ethnic groups of their generation. This suggests that (er) is no longer an ethnolectal marker for this generation.

We can also see that the Greek-Australian Young Adult women are matching the duration of their Anglo-Australian peers and the Greek-Australian women of the previous generation. The final model analysing just UWC and MC 2010s Young Adults did not have gender or community as a significant predictor (noting that all classes are included in Figure 3). Young Adult Greek-Australians are not significantly different from their Anglo- ($df = 21$, $t = 0.4$, $p = 0.7$) or Italian-Australian ($df = 20$, $t = -0.8$ $p = 0.43$) peers. There is also no interaction between community and gender; although the

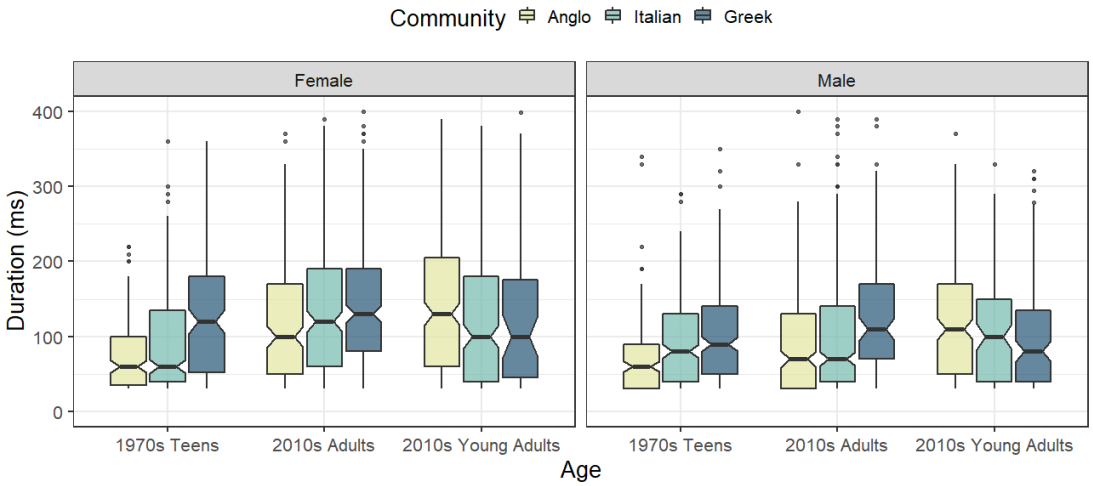


Figure 3 (er) *duration by age, gender, and ethnic community for the 1970s Teens, 2010s Adults and Young Adults*

Young Adult Greek-Australian men have shorter realisations of (er) relative to their male 2010s Young Adult peers, and Greek-Australian men of the previous generation, these differences are not significant.

Interestingly, while speech rate remains a significant predictor for the 2010s Young Adults, there is no longer a significant interaction with community. Speech rate therefore similarly affects (er) duration in this age group across communities, with faster speech rate still inhibiting duration.

Overall, it is evident that Greek-Australian Young Adults are not progressing (er) lengthening further than the previous generation of their community. They are patterning with their own generation in the 2010s. For this generation, (er) therefore is no longer marking Greek-Australian ethnicity as it was for the previous generation of 1970s Teens and 2010s Adults.

4. Discussion and Conclusions

This analysis of (er) sought to address an under-examined question about the longevity of ethnolectal differences within a speech community. It aimed to establish whether an ethnic group that led a community change forty years ago remained ahead of other groups today and, by extension, whether this variable continues to be an ethnolectal marker. The analysis addresses a specific question arising from the study by Grama, Travis and Gonzalez [4], due to the lack of data from Greek-Australian 2010s Adults and Young Adults. Examination of data from these groups here has allowed us to verify their hypothesis that the ethnolectal status of (er) has diminished over time as the lengthening led by the Greek-Australian community was taken up by other ethnic groups. While the data overall support this hypothesis and the original conclusions, focusing on two distinct stages of the change highlighted important generational distinctions in (er) realisation. The ethnolectal status of (er) is sustained for the 2010s Adults and diminishes specifically in the next generation of 2010s Young Adults.

When the Greek-Australian 2010s Adults and Young Adults are included in the analysis of (er), we observe that the 2010s Adults have continued to progress the change in much the same way as their Anglo- and Italian-Australian peers. We also observe that the Greek-Australians maintain their comparatively longer realisations of (er) over time within the generation of 1970s Teens and 2010s Adults. The Greek-Australian 2010s Adults remain significantly longer than the Anglo- and Italian-Australian 2010s Adults in the same way that the Greek-Australian Teens were significantly longer than the Anglo- and Italian-Australian 1970s Teens. This appears to be driven at least in part by Greek-Australian MC Adults lengthening their (er) duration in the 2010s to meet the benchmark established by their (L)WC peers in the 1970s. (er) therefore appears to be a stable ethnolectal marker of the Greek-Australian community for this generation. While the inhibiting effect of faster speech is more pronounced for the Greek-Australian 2010s Adults, their longer (er) duration is maintained regardless of this effect.

It is in the next generation that the ethnolectal status of (er) has shifted. There is no evidence that Greek- Australian Young Adults have continued to progress (er) lengthening beyond the benchmark set by the previous generation. The Greek-Australian Young Adults are not significantly longer than their Anglo-, or Italian-Australian peers. Instead, they are patterning with the other ethnic groups in their age group. This has taken place alongside reduced lengthening between the 2010s Adults and Young Adults, compared to between the 1970s and 2010s Adults, and a reduction in gender differences for the youngest age group. The collective evidence therefore indicates that the lengthening of (er) has begun to plateau in the Sydney speech community and no longer functions as an ethnolectal marker of the Greek-Australian community for 2010s Young Adults.

Through analysing the realisation of (er) duration in spontaneous speech across and within generations, this analysis has provided insight into the progression and plateau of a linguistic change led by a minority ethnic community. I propose that, consistent with the conclusions reached by Grama, Travis and Gonzalez [4], (er) is an example of a linguistic feature that functions both as an ethnolectal marker for members of the Greek-Australian community who were born in the 1960s, and as a feature that has been adopted by the wider speech community. As (er) lengthening is progressively taken up in the speech community its status as an ethnolectal marker diminishes, including for the ethnic group that originally utilised it most markedly as an ethnolectal feature.

The longevity of (er) as an ethnolectal marker of the Greek-Australian community appears to be limited to a single generation. This variable therefore highlights the importance of analysing ethnolectal variables in real- as well as apparent-time, as the social meaning of such variables may change along with the community over time.

5. Acknowledgments

I gratefully acknowledge the support I have received from the ARC Centre of Excellence for the Dynamics of Language, and the Sydney Speaks project team. I would like to acknowledge the previous analysis done on this variable by James Grama, Catherine Travis and Simon Gonzalez. I thank Catherine Travis, James Grama and Gan Qiao for their feedback, as well as the three anonymous reviewers for their comments on the paper.

6. References

[1] J. Cheshire, P. Kerswill, S. Fox, and E. Torgersen, "Contact, the feature pool and the speech community: The emergence of Multicultural London English," *Journal of Sociolinguistics,* vol. 15, no. 2, pp. 151-196, 2011.

[2] J. Gross, S. Boyd, T. Leinonen, and J. A. Walker, "A tale of two cities (and one vowel): Sociolinguistic variation in Swedish," *Language Variation and Change,* vol. 28, no. 2, pp. 225-247, 2016.

[3] B. M. Horvath, *Variation in Australian English: the sociolects of Sydney* (Cambridge studies in linguistics). Cambridge: Cambridge University Press, 1985.

[4] J. Grama, C. E. Travis, and S. Gonzalez, "Ethnolectal and community change ov(er) time: Word-final (er) in Australian English," *Australian Journal of Linguistics,* vol. 40, no. 3, pp. 346-368, 2020.

[5] M. Clyne, E. Eisikovits, and L. Tollfree, "Ethnic Varieties of Australian English," in *English in Australia*, D. Blair and P. Collins Eds. Amsterdan/Philadelphia: John Benjamins Publishing Company, 2001, pp. 223-239.

[6] J. Warren, "'Wogspeak': Transformations of Australian English," *Journal of Australian Studies,* vol. 23, pp. 85-94, 1999. [Online]. Available: http://www-tandfonline-com.ezproxy1.library.usyd.edu.au/doi/pdf/10.1080/14443059909387503?needAccess=true.

[7] S. Kiesling, "Variation, stance and style: Word-final -er, high rising tone, and ethnicity in Australian English," *English World-Wide,* vol. 26, no. 1, pp. 1-42, 2005. [Online]. Available: http://www.pitt.edu/~kiesling/kiesling-ER.pdf.

[8] J. Clothier, "Heading South: Phonetic differences in word finaler in speakers from Lebanese-and Anglo-Celtic Australian ethnic groups," presented at the Australian Linguistic Society Annual Conference, University of Newcastle, 12-12 December 2014, 2014.

[9] C. E. Travis, J. Grama, and S. Gonzalez. *Sydney Speaks Corpora*. [Online]. Available: http://www.dynamicsoflanguage.edu.au/sydney-speaks/

[10] H. Lausberg and H. Sloetjes, "Coding gestural behavior with the NEUROGES-ELAN system," *Behavior research methods,* vol. 41, no. 3, pp. 841-849, 2009.

[11] *LaBB-CAT*. (2012). New Zealand Institute of Language, Brain and Behaviour, University of Canterbury, NZ. [Online]. Available: http://labbcat.sourceforge.net/

[12] J. W. Du Bois, S. Schuetze-Coburn, S. Cumming, and D. Paolino, "Outline of discourse transcription," in *Talking data: Transcription and coding in discourse research*, J. A. Edwards and M. D. Lampert Eds. Hillsdale: Lawrence Erlbaum Associates, 1993, pp. 45-89.

[13] S. Gonzalez, J. Grama, and C. E. Travis, "Comparing the performance of forced aligners used in sociophonetic research," *Linguistics Vanguard,* vol. 6, no. 1, 2020.

[14] 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.

[15] *R: A language and environment for statistical computing*. (2020). R Foundation for Statistical Computing, Vienna, Austria. [Online]. Available: http://www.R-project.org/

[16] 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, 12/06 2017, doi: 10.18637/jss.v082.i13.

[17] *AICcmodavg: Model selection and multimodel inference based on (Q)AIC(c)*. (2020). [Online]. Available: https://cran.r-project.org/package=AICcmodavg

Ethnicity and social class in pre-vocalic *the* in Australian English

Gan Qiao, Catherine E. Travis

ARC Centre of Excellence for the Dynamics of Language, Australian National University

qiao.gan@anu.edu.au, catherine.travis@anu.edu.au

Abstract

Across varieties of English, the realisation of pre-vocalic definite article *the* is undergoing change, with a move away from non-glottalised FLEECE towards both glottalised FLEECE and SCHWA. Here, we examine this change in apparent time in a socially stratified corpus of spontaneous speech with 91 native speakers of Australian English of Anglo-Celtic, Italian and Chinese background. Analyses of 1,207 tokens reveal that the increase of both new forms is led by women and middle class speakers. An overall higher rate of newer variants by Italian and Chinese Australians is attributable not directly to ethnicity, but to its intersection with social class.

Index Terms: sociophonetics, sound change, *the*+VOWEL, ethnicity, social class

1 Introduction

Across most varieties of English, definite article *the* shows morphophonological alternation, occurring with a FLEECE vowel before another vowel (e.g. *the other* [ðiː ˈʊðə]), and SCHWA before a consonant (e.g. *the same* [ðə ˈsæɪm]) [1]. The pre-vocalic environment, however, is undergoing change, with an increase in SCHWA and in glottalisation found for American English [2], British English [3, 4], New Zealand English [5, 6], and, most recently, Australian English [7]. This change has been reported to be led by ethnic minorities and in ethnically diverse communities, in accordance with the understanding that greater diversity serves to drive change forward [3, 4, 6, 7].

Here, we examine change in apparent time in Australian English, as observed in a corpus of sociolinguistic interviews conducted in the 2010s in Sydney with Adults (born 1960s) and Young Adults (born 1990s). We consider two ethnic minority groups who have not been included in the *the*+VOWEL literature to date, Italian and Chinese Australians, and compare their patterning with that of Anglo-Celtic Australians. To better discern the nature of the change, we consider the increase in SCHWA and increase in glottalisation following FLEECE independently of each other, and find that, while these two changes are parallel in many ways, they are not identical.

Prior work on this variable has primarily focused on the increase of SCHWA relative to FLEECE over time [2-4, 6], and provides evidence that this is quite a recent change, with minimal SCHWA being reported for older speakers [6], or in older data [2], and with rates in some cases as high as 100% for younger speakers [3, 4, 6]. An increase in glottalisation with FLEECE has received less attention, but has also been observed [5], and in one case, found to be greater than the increase in SCHWA [7].

A consistent pattern reported for the social conditioning of this change has been the leading role of ethnically diverse speakers: Bangladeshi boys in London, UK [3, 4]; more diverse neighbourhoods in Auckland, New Zealand [6]; and Lebanese Australians in Sydney, Australia [7]. Beyond this, results for the social nature of these changes vary. In the UK, the increase in SCHWA has been found to be led by young men [3, 4]. In Australia, on the other hand, Cox and colleagues found that women were in the lead [7], and a similar tendency was observed by Meyerhoff and colleagues in New Zealand [6]. Social class has not been widely examined, though Hay and colleagues found "non-professional" speakers were ahead in the change in New Zealand [5]. These results thus do not present a clear picture of the social nature of the move away from non-glottalised FLEECE.

In the data examined here, we also find that the Italian and Chinese Australians are ahead of their Anglo-Celtic peers in the move away from non-glottalised FLEECE. To interpret this, we consider ethnicity alongside social class. Though most often considered independently, prior work has observed links between ethnicity and class, from Laferriere's classic study of ethnic variation in which she reported an effect of education on the differential uptake of changes by Jewish and Italian Americans [8] (cf. also [9]). Likewise, a favouring of standard features in British Asian English in contrast to Multicultural London English has been tied to the higher socioeconomic status of the former [10]. For *the*+VOWEL, we find that apparent ethnic differences are diminished once we take into account social class, leading us to conclude that the two must be considered together.

2 Methods

The data come from the Sydney Speaks corpus, a socio-linguistic corpus comprising recordings made in the 1970s-1980s and the 2010s with some 260 native speakers of Australian English [11]. In this paper, we report on the speech of 91 participants recorded from 2014 to the present. As summarised in Table 1, this includes Adults (born 1960s) and Young Adults (born 1990s) of Anglo-Celtic background and Young Adults from two migrant communities, Italian and Chinese Australians, all of whom were born and raised in Australia (or arrived before the age of six) to parents from Italy and Hong Kong or Guangzhou, China. There are even numbers of men and women for each community group.

The sample is further stratified by social class, determined on the basis of a composite measure of occupation, level of education and school type, and broken into three groups: Middle, Lower-Middle and Working Class. All three class groups are represented for Anglo-Celtic and Italian Australians, though there is a predominance of the Lower-Middle Class for the latter, due to the fact that, at the current stage of data collection, one half of the Italian Australians are university students. For the Chinese Australians, there are no Working Class participants, a reflection of the generally high socio-economic status of this community (as evident from census reports [12], and as described in [13]). This social distribution turns out to be key to interpreting the patterns observed.

Table 1. Demographic breakdown of participants.

	Adults (45-64 yrs)	Young Adults (18-32 yrs)		
	Anglo	Anglo	Italian	Chinese
Middle	9	9	5	9
Lower-Middle	10	10	9	13
Working	8	7	2	

Sociolinguistic interviews were conducted by community members (Anglo-Celtic, Italian and Chinese Australians from Sydney), primarily within their own networks, with friends, extended family, and other acquaintances. They lasted between 60 and 90 minutes, and approximately 30 minutes (or 5,000 words) were transcribed per participant, providing a total of over 500,000 words for the analyses presented here.

All instances of *the* preceding a vowel-initial word were extracted [using the search function in LaBB-CAT, 14], giving an initial total of 1,708 tokens. We set aside tokens occurring in overlap or where there was background noise and those where *the* was not immediately followed by another word (including those followed by a pause, a filled pause such as *um,* or truncation) (n = 431). All remaining tokens were auditorily coded blind by two trained listeners, with spectrograms also reviewed in Praat in some cases for clarification. Tokens were coded both for realisation of the vowel in pre-vocalic *the* (vowel type, FLEECE vs. SCHWA) and for glottalisation following the vowel (present vs. absent, without distinguishing type or duration of glottalisation, following [7]). Coders also noted cases where they were unable to make a determination, including where the vowel in *the* was assimilated with the following vowel (n = 41), and instances of emphatic *the* (n = 2, both of which were realised as FLEECE with no glottalisation).

All discrepancies between the coders were checked by a third coder, auditorily for vowel type (n = 152) and by reviewing the spectrograms in Praat for glottalisation (n = 90). In most cases, this allowed for a determination to be made, though a small number of cases were unidentifiable, generally due to assimilation with the following vowel, and excluded (n = 27). This left a total of 1,207 tokens which had been coded both for vowel type and glottalisation.

Of the four possible realisations, non-glottalised FLEECE (the older form) is the majority variant, accounting for close to two-thirds of all instances (n = 776). The remaining tokens are largely made up of glottalised FLEECE (13%, n = 155) and glottalised SCHWA (20%, n = 243), as non-glottalised SCHWA is vanishingly rare (n = 33), as also reported in other studies [2, 4, 5, 7]. We, therefore, set aside non-glottalised SCHWA in the analyses that follow, leaving a total of 1,174 tokens for analysis distributed across three variants.

Figure 1 presents spectrographic representations of each of these three variants. Non-glottalised FLEECE appears on the left where no stop closure between *the* and the following vowel is evident, in contrast with the irregular glottal vibrations seen in the glottalised realisations in the middle (also for FLEECE) and on the right (for glottalised SCHWA, which we will simply refer to as SCHWA from here on).

3 Results

3.1 An overview of *the* alternations

We begin by considering the overall distribution of the three variants across age and ethnicity, to gain an overview of the change in apparent time for the Anglo Australians (comparing the Adults and Young Adults), and of community differences for the Young Adults (comparing the Anglo, Italian and Chinese Australians). Figure 2 presents the rates of the three variants for Adult Anglos in the first set of columns, then Young Adult Anglos, Italians and Chinese respectively.

This chart indicates that, for the Adult Anglos, FLEECE with no glottalisation (at a rate of 83%) is used nearly to the exclusion of the other two variants, neither of which reaches 10%. For the Young Adult Anglos, the proportion of FLEECE with no glottalisation has dropped substantially (to 64%), while both FLEECE and SCHWA with glottalisation have increased (to 21% and 16% respectively). Thus, in these speakers, we see the same apparent time change that has been reported elsewhere.

Also consistent with other reports, the ethnic minorities appear to be ahead in this change – both the Italian and Chinese Australians exhibit a lower rate of non-glottalised FLEECE than their Anglo peers (at 56% and 47% respectively), and a higher rate of SCHWA (at 31% and 37%). Glottalised FLEECE, however, is produced at a similar rate across all three groups.

This differential patterning for vowel type suggests that the change is twofold, involving an increase in glottalised FLEECE that applies across all three communities equally, as well as an increase in SCHWA, in which the ethnic minorities are in the lead. Confirming the independence of the two incoming variants is the fact that speakers' rates of use of these forms relative to the older non-glottalised FLEECE do not correlate; that is, a high rate of SCHWA is not indicative of a correspondingly high rate of glottalised FLEECE, or vice versa (for the 72 speakers who produce at least five tokens of glottalised and non-glottalised FLEECE, and at least five tokens of SCHWA and non-glottalised FLEECE, $r(70) = .17$, $p = .146$). We will see below that it is not only in the rate of uptake by the ethnic minorities that these two incoming forms differ.

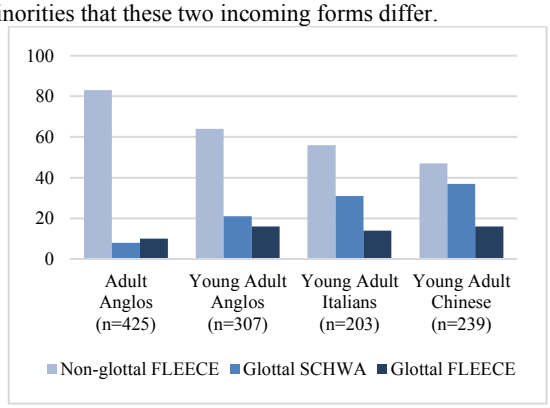
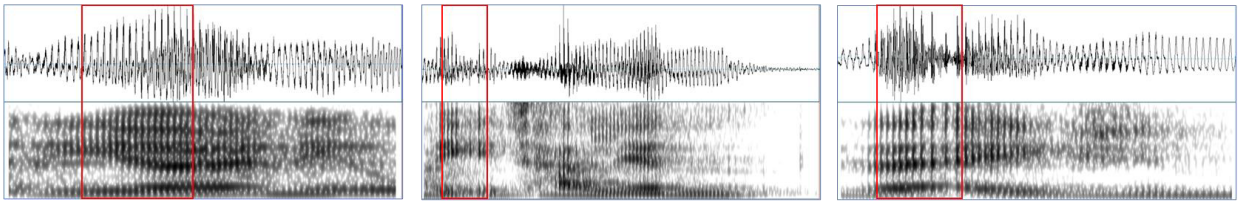


Figure 2: *the*+VOWEL realisations for Age by Community



Non-glottalised FLEECE: *the only* | Glottalised FLEECE: *the Italian* | Glottalised SCHWA: *the Asian*

Figure 1: *Spectrographic representations of three variants for the+VOWEL.*

3.2 Modelling the variation

In order to model the variation, we analysed separately the alternation between glottalised and non-glottalised FLEECE on the one hand, and SCHWA and non-glottalised FLEECE on the other. We also analysed separately the patterning for the Anglo Australians, comparing the Adults and Young Adults to assess change in apparent time among the majority community, and the patterning for the Young Adults, comparing the Anglo, Italian and Chinese Australians, to assess the role of ethnicity in the change. Thus, four independent analyses were conducted, and are reported below. Analyses were done using mixed-effects logistic regression, with the lme4 package in R [15, 16].

For each analysis, realisation was the dependent variable, with the more conservative, non-glottalised FLEECE, the reference level. The social predictors of Socio-economic Class and Sex were included in all models, and Age was included in the model for Anglo Australians (Section 3.2.1), and Ethnicity in that for the Young Adults (Section 3.2.2).

For linguistic predictors, we included Stress and Height of the word-initial vowel in the following word, based on reports in the literature of a favouring of SCHWA when the following syllable is stressed [17, 18], and with high-front vowels (in particular, FLEECE), in what has been interpreted as a dissimilation effect [6, 17]. Observation of the patterns evident in the raw data indicated that the most meaningful classification for the following vowel was a three-way Height distinction between High, Mid and Low vowels. 80% or more of each of these categories was made up of a small set of vowel types, respectively KIT and FLEECE (the most frequent forms being *Italian, eastern*); GOAT, DRESS, and LOT (e.g. *only, end, Australian)*; and STRUT, TRAP, and FACE (e.g. *other, accent, Asian*). And finally, though not included in the previous literature, we tested for an effect of Word Class, distinguishing between Adjectival and Nominal uses (e.g. *other, only* vs. *end, area*) determined by considering each instance in context.

Speaker and Word were included as random intercepts. All word types that occurred only once (approximately one half of all word types) were pooled into a single level as, with such low frequency, they would not have lexically specific probabilities, and their inclusion may make the model oversensitive to lexical particularities [19-21]. We tested for relevant three- and two-way interactions, and then pruned the models to leave only significant predictors.

We now turn to these analyses, first presenting the apparent time comparison among Anglo Australians, before moving on to consider ethnic variation for the Young Adults.

3.2.1 The+VOWEL in apparent time

We have already noted the minimal use of both glottalised FLEECE and SCHWA for the Anglo Adults, and the substantial increase for both forms for the Young Adults (Figure 2 above). The regression analyses confirm the increase in these variants in apparent time, with the Young Adults having a significantly higher rate of each than the Adults (as described below). However, comparisons of the two regression analyses reveal both similarities and differences in the way in which glottalised FLEECE and SCHWA are coming into the system.

Both newer variants are most favoured with women (for glottalised FLEECE, $\beta = 0.81$, $p < .05$ and for SCHWA, $\beta = 1.29$, $p < .05$), and with the Middle Class (see below), and they are both favoured with high stressed vowels (e.g. with *eastern, inner, English*). But they differ, first, in that glottalised FLEECE appears to be more sensitive to linguistic conditioning than SCHWA. Glottalised FLEECE is favoured by Adjectives over Nouns ($\beta = 0.73$, $p < .05$), but there is no Word Class effect for SCHWA. There is also a marginal main effect of vowel Height for glottalised FLEECE, which is most favoured with High vowels (vs. Low vowels, $\beta = 0.96$, $p = 0.074$), while for SCHWA, an effect of vowel Height exists only in interaction with Stress. There are also some differences in the social conditioning, as presented in Figure 3 for glottalised vs. non-glottalised FLEECE, and Figure 4 for SCHWA vs. non-glottalised FLEECE. Glottalised FLEECE is favoured by the Young Adults over the Adults ($\beta = 0.78$, $p < .05$), and, for both age groups (evidenced in the lack of an interaction), by the Middle Class over both the Lower-Middle ($\beta = -0.88$, $p < .05$) and Working Class ($\beta = -0.79$, $p = .095$).

For SCHWA, Age has a significant main effect, such that SCHWA is favoured by the Young Adults over the Adults ($\beta = 3.44$, $p < .001$), but there is also a significant interaction with Social Class. While there is no Class effect for the Adults, for the Young Adults, SCHWA is most favoured by the Middle Class. This interaction suggests that class differentiation is more recent for SCHWA than for glottalised FLEECE. It also appears to be more defined for SCHWA, which exhibits a gradual weakening in the favouring effect from Middle to Lower-Middle to Working Class, while for glottalised FLEECE, the Middle Class has the highest rate, but there is no difference between the Lower-Middle and Working Class. We now turn to consider how these differences surface when we compare the different communities for the Young Adults.

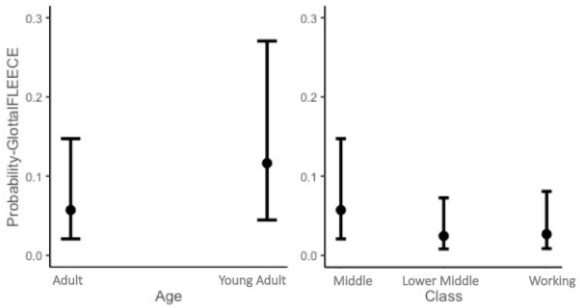


Figure 3: Predicted probability of glottalised FLEECE vs. non-glottalised FLEECE for Age (left) and Class (right): Anglo Australians

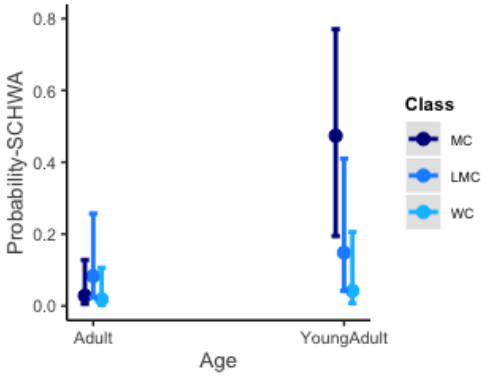


Figure 4: Predicted probability of SCHWA vs. non-glottalised FLEECE for Age by Class: Anglo Australians

3.2.2 The+VOWEL across ethnicity

As we saw above in Figure 2, the rate of glottalised FLEECE is similar across communities, while for SCHWA, both the ethnic minorities have a higher rate than the Anglo Australians.

Before concluding that ethnic minorities are leading in the change to SCHWA, it is important to bear in mind the patterning for Social Class among the Anglo Australians that we have just seen, specifically, the lagging of the Working Class in this change. The fact that there is a higher proportion of Working Class Anglos in the sample than both Italians (due to sampling) and Chinese (reflective of the nature of the community; Table 1) means that the samples are not directly comparable.

In order to meaningfully compare the patterning of the Young Adults, we, therefore, set aside the Working Class participants, and ran regression analyses on the 55 remaining participants, making a two-way distinction between Middle and Lower-Middle Class. We ran the same two analyses that we ran for the Anglo Australians, comparing glottalised and non-glottalised FLEECE, and then SCHWA and non-glottalised FLEECE.

For the linguistic predictors, we find generally similar effects across the communities as those we reported above: a favouring of glottalised FLEECE with High vowels (as a main effect vs. Low vowels, $\beta = 1.66$, $p < .05$), and in interaction with Stress, and with Adjectives ($\beta = 0.83$, $p < .05$); and a favouring of SCHWA with Stress ($\beta = -4.39$, $p < .0001$), with no effect for vowel Height or Word Class. There are no significant interactions between Community and the linguistic predictors, indicating that all groups draw on the same linguistic system in participating in this change. What of the social effects?

For glottalised FLEECE, there is no effect for Community, neither as a main effect nor in interaction with Sex or Social Class. For all communities, glottalised FLEECE is favoured by women ($\beta = 0.66$, $p < .05$) and there is no significant effect for Social Class. This is consistent with the parallel rates seen across the communities in Figure 2.

For SCHWA also, across all communities, there is a favouring effect by women ($\beta = 1.72$, $p < .01$), but, crucially, there is no main effect for Community, despite the higher rates by Italian and Chinese Australians (Figure 2). The Community effect emerges only in interaction with Social Class. As depicted in Figure 5, for the Anglo Australians, the Middle Class favours SCHWA over the Lower-Middle Class, but for the Italian and Chinese Australians, there is no significant difference between the two social classes (as verified by separate analyses). For the Chinese, both class groups pattern similarly to the Middle Class Anglo Australians, and for the Italians, both lie in between the Middle and Lower-Middle Class Anglo Australians. It is this differential class patterning, then, that boosts the rate of SCHWA overall for the Italian and Chinese Australians.

Why might there be greater ethnic difference for SCHWA than for glottalisation of FLEECE? This can partly be accounted for by the weaker stratification overall, seen in Figure 4, with the Lower-Middle Class patterning similarly to the Working Class. Further, it may also be partly due to the overall lower rate of occurrence of this variant, which minimises the impact of the class distribution.

4 Discussion and conclusions

The analyses of spontaneous speech in Sydney corroborate findings across the English-speaking world, including in Australia, that there is an ongoing move away from non-glottalised FLEECE in *the*+VOWEL. This change is being led by young women, in accordance with previous findings in Australia [7], but distinct from what has been observed in the UK [3, 4]. Here, we have also identified a leading role for the Middle Class speakers, and this, coupled with the sex effect, may be indicative of the social meaning of these newer variants in Australia, and an association with overt prestige.

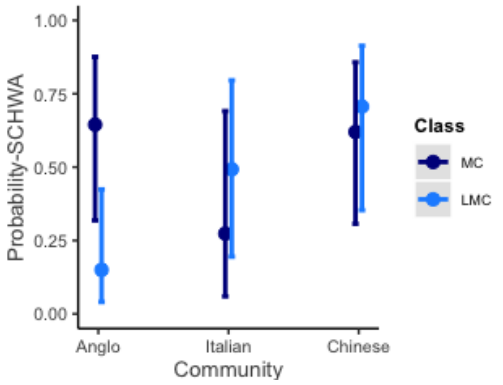


Figure 5: Predicted probability of SCHWA vs. non-glottalised FLEECE, for Community by Social Class: Young Adults.

As in other studies, the shift away from non-glottalised FLEECE is greatest for ethnic minorities, here, applicable specifically to an increase in SCHWA, while glottalised FLEECE is parallel across the three communities. The explanation for this apparent community difference lies in the intersection between social class and ethnicity. First, the rate of SCHWA for the Anglo Australians is driven down by the higher proportion of Working Class speakers, who disfavour this variant. Second, a favouring of SCHWA by Middle over Lower-Middle Class Anglo Australians is not replicated for Italian and Chinese Australians, for whom there is no distinction between these class groups. Furthermore, both the Middle and the Lower-Middle Class Italian and Chinese Australians exhibit greater favouring of SCHWA than the Lower-Middle Class Anglo Australians, thus bumping up the overall rate for these communities.

General patterning in accordance with middle-class norms has been observed for Chinese Australians across a range of variables, for which they appear to strongly favour forms that are associated with overt prestige, including, for changes in diphthong realisations [22]; the pre-nasal vs. pre-obstruent split for TRAP [23]; velar vs. alveolar realisations of (ing) [24]; and, beyond phonetics, the long-term change from *will* to *be going to* [25]. Given the generally high socio-economic status of this community, it appears that part of being a Chinese Australian is belonging to the Middle Class, with speech patterns being one mechanism by which that may be marked. The Italian community overall is not afforded the same social status, but in the sample here, eight of the nine Lower-Middle Class participants are university students, who may be experiencing upward mobility, rendering them potentially more attentive to norms of overt prestige.

As this is a change driven by the Middle Class, this puts the Italian and Chinese Australians examined here at the forefront of the change. However, they lag in changes that are not led by the Middle Class [23, 25]. Thus, their role as leaders we would argue is not a direct artefact of greater diversity, but of the "particular esteemed educational, occupational, and linguistic values" [8] of these communities in combination with the social meaning of the variable under consideration.

We conclude that apparent ethnic differences should be assessed in relation to the broader social differentiation of the relevant variables, and general social contextualisation of the ethnic groups under study, as this will allow us to better understand the ways in which ethnic diversity may impact patterns of language variation and change.

5 Acknowledgements

The Sydney Speaks project has been developed with support from the ARC Centre of Excellence for the Dynamics of Language. We gratefully acknowledge the work of the Sydney Speaks project team in compiling the corpus, and also thank Benjamin Purser, Thomas Powell-Davies and Thea Shillam for assistance with the coding, and three SST anonymous reviewers for valuable feedback on an earlier version of this paper.

6 References

[1] Quirk, R., Greenbaum, S., Leech, G., and Svartvik, J., A comprehensive grammar of the English language, London: Longman, 1985.

[2] Todaka, Y., "Phonetic variants of the determiner "the"," UCLA Working Papers in Phonetics, 81:39-47, 1992.

[3] Cheshire, J., Kerswill, P., Fox, S., and Torgersen, E., "Contact, the feature pool and the speech community: The emergence of Multicultural London English," Journal of Sociolinguistics, 15(2):151-196, 2011.

[4] Fox, S., The New Cockney, New York: Palgrave Macmillan, 2015.

[5] Hay, J., Walker, A., McKenzie, J., and Nielsen, D., "The changing realisation of 'the' before vowels in New Zealand English," New Zealand English Journal,, 26:23-32, 2012.

[6] Meyerhoff, M. et al., "Definite change taking place: Determiner realization in multiethnic communities in New Zealand," University of Pennsylvania Working Papers in Linguistics (Selected Papers from NWAV 47), 25(2):71-78, 2020.

[7] 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:1-31, 2022.

[8] Laferriere, M., "Ethnicity in phonological variation and change," Language, 55(3):603-617, 1979.

[9] Boberg, C., "Real and apparent time in language change: Late adoption of changes in Montreal French," American Speech, 79(3):250-269, 2004.

[10] Sharma, D., "Prestige factors in contact-induced grammatical change," Advancing Socio-Grammatical Variation and Change: In Honour of Jenny Cheshire, Beaman, K. V. et al., eds., 55-72, Oxon/New York: Routledge, 2021.

[11] Travis, C. E., Grama, J., and Gonzalez, S., "*Sydney Speaks Corpora*," http://www.dynamicsoflanguage.edu.au/sydney-speaks/, In Progress.

[12] Australian Bureau of Statistics. "Census of Population and Housing," 25 Sept, 2015; https://www.abs.gov.au/AUSSTATS/abs@.nsf/DetailsPage/2101.01991.

[13] Grama, J., Travis, C. E., 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.

[14] Fromont, R., and Hay, J., "LaBB-CAT: An annotation store," Proceedings of the Australasian Language Technology Workshop:113-117, 2012.

[15] Bates, D. et al., "lme4: Linear mixed-effects models using 'Eigen' and S4. R package, version 1.1-121," 2019.

[16] R Core Team, R: A language and environment for statistical computing, Vienna, Austria: R Foundation for Statistical Computing, http://www.R-project.org 2019.

[17] Gaskell, M. G. et al., "Constraints on definite article alternation in speech production: To "thee" or not to "thee"?," Memory & Cognition, 31(5):715-727, 2003.

[18] Raymond, W. D., Healy, A. F., McDonnel, S., and Healy, C. A., "Acquisition of morphological variation: The case of the English definite article," Language and Cognitive Processes, 24(1):89-119, 2009.

[19] Barth, D., and Kapatsinski, V., "Evaluating linguist mixed-effects models of corpus-linguistic data in light of lexical diffusion," Mixed-effects regression models in linguistics: Quantitative methods in the Humanities and Social Sciences, Speelman, D. et al., eds., 99-116, Cham: Springer, 2018.

[20] Szmrecsanyi, B., Biber, D., Egbert, J., and Franco, K., "Toward more accountability: Modeling ternary genitive variation in Late Modern English," Language Variation and Change, 28(1):1-29, 2016.

[21] Travis, C. E., and Torres Cacoullos, R., "Categories and frequency: Cognition verbs in Spanish subject expression," Languages (Special Issue 'Revisiting language variation and change: Looking at metalinguistic categories through a usage-based lens', Eds. Brown & Rivas), 6(3):126, 2021.

[22] Grama, J., Travis, C. E., and Gonzalez, S., "Ethnic variation in real time: Change in Australian English diphthongs," Studies in Language Variation, Van de Velde, H. et al., eds., 292-314, Amsterdam: John Benjamins, 2021.

[23] Sheard, E., and Qiao, G., "Class, gender and ethnic differentiation of pre-nasal TRAP by young Chinese Australians," Paper presented at the Australian Linguistic Society Annual Conference, 2021, La Trobe University. 2021.

[24] Travis, C. E., Purser, B., and Grama, J., "Stability and change in (ing) over 40 years in Australian English: Ethnicity, gender and social class," Paper presented at the Australian Linguistic Society Annual Conference, 2021, La Trobe University. 2021.

[25] Qiao, G., "A comparative study of ethnic variation and change in Australian English," PhD, Linguistics, Australian National University, In Progress (2019-).

Variation in /t/ in Aboriginal and Mainstream Australian Englishes

Debbie Loakes[1], Kirsty McDougall[2] and Adele Gregory[1]

[1]The University of Melbourne; [2]University of Cambridge

dloakes@unimelb.edu.au; kem37@cam.ac.uk; adele.gregory@unimelb.edu.au

Abstract

This paper describes regional and social variation in /t/ in Australian Englishes (Aboriginal and Mainstream). Data are from two dialect communities and two regions in Victoria, Warrnambool and Mildura. Results show strong sociophonetic patterning for dialect and region, as well as age-related variability. Some patterns also emerge for gender for certain /t/ variants. This paper describes a possible link with consonant variability and voice quality; speaker groups that produce stops with more glottal constriction have been shown in other studies to have more laryngeal activity in vowels.

Index Terms: Australian Aboriginal English, Mainstream Australian English, voice quality, t-quality, regional variation

1. Introduction and Background.

1.1. /t/ variation in Mainstream Australian English

While early work on sociophonetic variation in Mainstream Australian English (MAE) tended to focus on vowels, research on other type of speech variability, including consonant realisation, has been growing in recent years. Production of voiceless plosives, especially /t/ has been demonstrated to yield extensive sociophonetic variation in a range of regions and speaking situations in this variety.

One of the first studies of MAE /t/ was conducted in Sydney [1] finding flap (tap), affricated, and aspirated variants through an auditory analysis of medial /t/. Patterns of usage also varied with sociolinguistic factors such as ethnicity and gender [2]. Auditory analysis of MAE /t/ in Sydney was also performed for two socioeconomic groups [2]. Occurrences of canonical /t/, glottal stop and tap variants were observed as well as a category labelled [ts] which combined heavily aspirated /t/, affricated variants and fully fricated variants. Variation was found according to gender and socioeconomic group. A further study of MAE English /t/ in Sydney [3] looked at /hVt/ citation words, finding /t/ is mostly produced with varying levels of glottalisation, with a small number of fricated and unreleased tokens observed. In a later study focusing specifically on glottalisation of post-vocalic /t/ [4] noted age-grading in production of this emerging variant, but not perception.

In Queensland, production of MAE /t/ by Brisbane school children was examined auditorily [5]. That analysis noted flap, glottalised, unaspirated, unreleased (sometimes with glottalisation), deleted, and aspirated variants. Sociophonetic variation was also observed, with unreleased /t/ used more frequently by children at fee-paying versus state schools.

Variation in MAE /t/ in Victoria has also been the subject of several studies focusing on MAE. An auditory study of speakers from Melbourne and surrounding rural areas from two socioeconomic groups [6] found four variants: canonical /t/ and tapped /t/ for both socioeconomic groups, fricated /t/ which was more popular in older speakers and those in the higher socioeconomic group, and glottalised /t/ which patterned differently with context both groups. Another study of word-medial and word-final /t/ by female university-educated speakers from Melbourne [7] found most realisations in that sample were fricated, but with some canonical, 'intermediate' and tap variants also observed. Intermediate tokens are described as having a fricative-like nature, sounding like fricatives, but with burst characteristics evident in the spectrum [7]. Meanwhile [8] studied the spontaneous speech of male twins from Melbourne and observed canonical, intermediate, tapped, aspirated and glottal variants and only a few fricated realisations. In a study by [9], spontaneous speech produced by primary school children from Yarrawonga (near the Victorian-NSW border) included released, unreleased, fricated, affricated, preaspirated, glottalised (glottal and laryngealised) and tapped variants of /t/. Gender-related patterns were observed, with greater use of pre-aspiration and fricated /t/ by females. Finally, in Western Australia, [10] looked at variability in conversational speech in Perth, finding /t/ patterned with speakers' neighbourhood of residence, the lexical item and "idiolect". Variants they observed [10:8] are "voiced/voiceless, canonical (released) plosive, fricative, tap, approximant, or absent".

1.2. /t/ variation in Aboriginal Australian Englishes

The research base on the phonetic characteristics of Aboriginal Australian Englishes (AAEs) is much smaller than for MAE, both generally and regarding /t/ variation specifically.

Broad overviews of AAE phonology, on the basis of impressionistic description, are provided by [11,12,13]. The sources of their data are not described in detail, but appear to be focused on L2 varieties (cf. [14]). The authors note the lack of a voiced/voiceless distinction for stops, while [11] elaborates that initial stops in AAE are typically voiced and unaspirated but also subject to wide intra-speaker variability. Intervocalic alveolar stops are described as often being produced with a tap/flap, similar to the mainstream variety.

More recently [15] provides a detailed acoustic analysis of stop production in three varieties of AAE (inc. L1 AAE) spoken on Croker Island, Northern Australia, in comparison with MAE spoken in Sydney. Measures of Voice Onset Time (VOT) and Voice Termination Time (VTT) were taken for /p t k b d g/ in read speech. Results showed that a voicing distinction was present for these varieties of AAE, with no significant differences present between AAE and MAE in the patterning of VOT across voicing categories. Differences between AAE and MAE were found in VTT, however, with the voiceless categories (including in L1 AAE) showing passive phonetic voicing. The authors surmise that this difference may be a factor that contributes to AAEs being perceptually different from MAE. The study also found considerable variability across L1 and L2 AAE stops, greater than that observed for the MAE

speakers./t/ in L1 AAE spoken in Warrnambool, Victoria, was analysed by [16], alongside MAE spoken in the same town. Acoustic profiling of read speech tokens in /hVt/ and /hVtV/ contexts revealed clear sociophonetic patterning across the speaker groups as well as within-group variation. The authors hypothesised different patterning across speaker groups may be due to a connection between voice quality and glottal timing, consistent with [17: 85] who note that segmental contrasts provide a "testbed" for the analysis of voice quality.

1.3. Aims

The aim of this study is to analyse sociophonetic variation in Australian English /t/, specifically:

1. to analyse whether there is regional variation in Australian English /t/ spoken in Victoria;
2. to determine whether previously observed variability in MAE and AAE holds in a different location (is dialect variation robust across locations?); and
3. to determine whether age- and gender-related differences exist in these communities.

This study also considers patterns in /t/ distribution with an existing analysis of voice quality for the same speakers [18].

2. Method and Analysis

2.1. Participants and experimental task

This study compares two groups of adult L1 English speakers from two locations in Victoria, Warrnambool (WN, a regional coastal city located in the southwest of Victoria) and Mildura (MI, a regional inland city located in the border region of the northwest of Victoria). Warrnambool is approximately 250km from Melbourne, while Mildura is approximately 550km from Melbourne (and around 400 km from Adelaide). The participants are 52 Australian English speakers: 24 AAE speakers (10M, 14F) and 28 MAE speakers (12M, 16F). By region, the breakdown of speakers is: MAE – WN 8 M, 7 F; MI 4 M, 9 F. The participants were all adults, identified as one of the binary gender categories male or female, and roughly fell into two equal age groups of <40 (18- 39) and >40 (40-72). Aboriginal participants identified themselves as being "Koori" or "Aboriginal". The data used in the current study was word list /hVt/ and /hVtV/ words, forming part of a larger study where participants also took part in a questionnaire, perception study and sociologuistic interview. In the present study, 2052 tokens were analysed overall (1209 tokens MAE, 843 AAE), an average of 39 tokens per speaker. Most tokens were word-final, while 201 tokens were /hVtV/.

2.2. Analysis

2.2.1. Phonetic analysis

Speech data were labelled using *Praat* [19] after autosegmentation of the phonemes using MAUS [20]. The overall quality of each /t/ (and release where present) was categorised auditorily and visually from spectrograms and annotated on a "phonetic" tier. A tier "t-category" recorded classification decisions. The following list slightly modified from [16] gives the category names and explanations about the decisions made during the labelling process. Further reference to the literature in determining these categories is made in [16]. Abbreviations (used in figures) are listed here.

Canonical [tʰ] (labelled C on figures): period of full closure followed by burst. No voicing apparent.
Affricate [tˢ] (labelled AF on figures): a closure followed by /s/-like release (not aspirated), no burst like characteristics.
Fricative [t̪] (F): a fully fricated variant, not the same as [s], better described as a "lowered /t/".
Intermediate (I): this category is best described as [t̪ʰ]. It has the auditory percept of a fricated stop, but there are burst characteristics evident acoustically.
Tap [ɾ] (T): durationally very short, only observed inter-vocalically.
Approximant [ɻ] (A) : technically a tap which does not have full closure, observed intervocalically.
Pre-glottalised [ʔ͡t] (PG): these stops have glottal activity and unreleased supralaryngeal closure.
Glottal [ʔ] (G): full glottal stopswith no apparent supralaryngeal closure characteristics. These stops can be either plosive-like or creaky in appearance; both were observed in the present data.
Ejective [t'] (E): Acoustically, ejectives tend to pattern in two ways: 1) with a period of closure followed by release of the supralaryngeal gesture, a period of "silence", and a second release which coincides with glottal opening, or 2) cases without the silence, where glottal opening occurs immediately after oral release. In the present data ejectives often show sharp "spikes" on the waveform correlated with burst intensity. Sometimes, two bursts are evident, one indicating release of supralaryngeal closure and another release of laryngeal closure.
Voiced [d̬] (D): these tokens are partially voiced, similar to what [15] describe for Aboriginal English groups, with passive phonetic voicing in what is a phonologically voiceless category.

2.2.2. Statistical analysis

In the analyses below data are analysed using both descriptive and inferential statistics. Statistical analysis was carried out using R and the rstatix package [21].

3. Results

3.1. Regional Variation

Data from all speakers were combined to determine if regional variation was present. A two-tailed, chi-square test of independence with a Bonferroni correction showed that regional variation was evident for *most* /t/ categories. For affricates WN speakers used significantly more affricates X^2 (1, 2,502) = 21.27, p<.001, fricatives X^2 (1, 2,502) = 19.18, p<.001, ejectives X^2 (1, 2,502) = 8.43, p<.05 and intermediate tokens X^2 (1, 2,502) = 54.98, p<.001. MI speakers, on the other hand, used significantly more canonical stops X^2 (1, 2,502) = 84.00, p<.001, and taps X^2 (1, 2,502) = 13.46, p<.01. No regional variability was evident for approximants, glottal stops and pre-glottalised tokens. For approximants, this is due to very minimal observations.
Looking more closely at the distributions in the data, in MI, over half of all /t/ were canonical stops, over 20% were affricates, over 10% were pre-glottalied tokens, and the remainder were split across the other variants (although no tokens were "intermediate" unlike in WN). In WN, as described in [16] canonicals and affricates were both observed at rates just over 30%, while fricatives and pre-glottalised stops were observed at a rate of around 10% each. As mentioned, this regional analysis includes all speakers, so it is important to look more closely at the sociophonetic patterning in each region.

3.2. Dialect variation, AAE vs MAE

The two dialects differed significantly in their use of t-categories for every variant except approximants. Further, AAE speakers were much more variable than MAE speakers. MAE speakers used significantly more affricates X^2 (1, 2,502) = 112.49, p<.001, canonical stops X^2 (1, 2,502) = 10.39, p<.001, fricatives X^2 (1, 2,502) = 59.29, p<.001 and intermediate tokens X^2 (1, 2,502) = 8.8, p<.05. AAE speakers used significantly more ejectives X^2 (1, 2,502) = 108.52, p<.001, pre-glottalised stops X^2 (1, 2,502) = 198.48, p<.001, glottal stops X^2 (1, 2,502) = 27.76, p<.001 and taps X^2 (1, 2,502) = 37.77, p<.001.
The patterning of /t/ variants across the two dialects can be seen more clearly in Figure 1, which shows the proportions of all /t/-categories used by AAE speakers (to the left of the figure, in black) and by MAE speakers (right, in yellow).

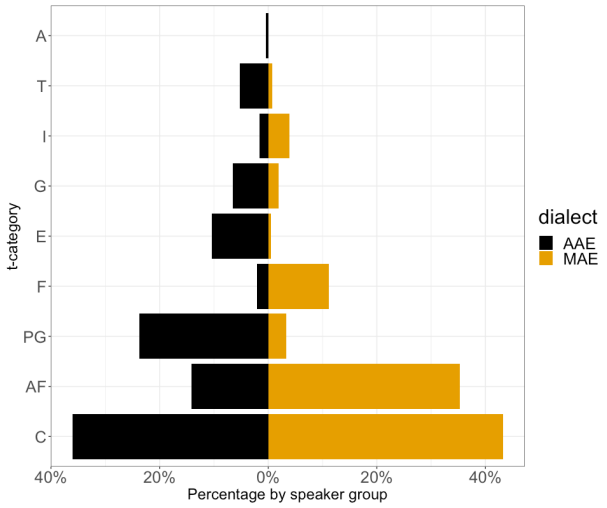


Figure 1: *back-to-back plot /t/ of variants in MAE and AAE*

It is also evident here that both dialects used all variants (except approximants which occurred in small numbers in the MAE group) but the distributions of variants are unequal. To give some examples, 23% of tokens produced by AAE speakers were pre-glottalised, in comparison with 3% for MAE. Affricates, on the other hand, occurred at a far higher rate for MAE speakers who used 35% overall compared to 14% tokens by AAE speakers.

3.3. Variability according to age

Variation according to age (whether under or over 40) was also significant for most variants. Speakers under 40 used more pre-glottalised variants X^2 (1, 2,502) = 115.43, p<.001, ejectives X^2 (1, 2,502) = 34.66, p<.001, glottal stops X^2 (1, 2,502) = 38.86, p<.001 and taps X^2 (1, 2,502) = 31.86, p<.001, while for participants over 40, there were more canonical stops X^2 (1, 2,502) = 12.62, p<.01, affricates X^2 (1, 2,502) = 72.63, p<.001 and fricatives X^2 (1, 2,502) = 24.59, p<.001. The most marked differences across the age groups were evident for glottal stops, pre-glottalised stops and ejectives (335 tokens), which were almost absent in the speech of older speakers (26 tokens).

3.4. Variability according to gender

/t/-categories also patterned with gender, but for a smaller number of variants and for the most part not as strongly (note again that the participants identified only as male or female). Women used significantly more canonical stops X^2 (1, 2,502) = 64.07, p<.001, while men used significantly more ejectives X^2 (1, 2,502) = 14.73, p<.01, intermediate /t/ X^2 (1, 2,502) = 8.96, p<.05, pre-glottalised stops X^2 (1, 2,502) = 18.22, p<.001 and taps X^2 (1, 2,502) = 10.20, p<.05. This patterning can be observed in Figure 2, which shows the proportions of all /t/-categories used by males (left , in yellow) and by females (right, in black).

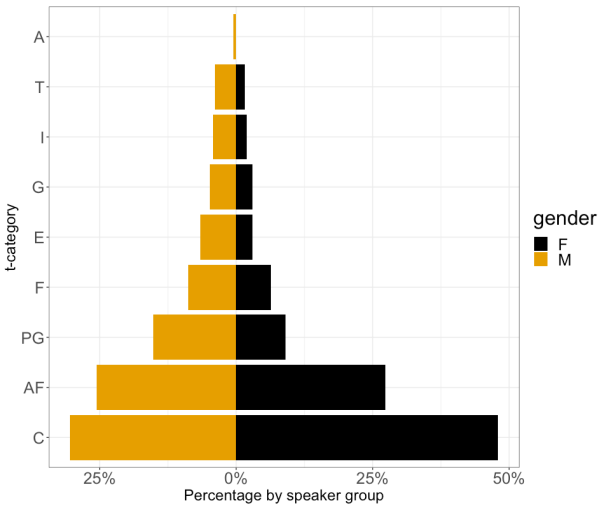


Figure 2: *back-to-back plot of /t/ variants by gender*

3.5. Broader patterns of /t/ variation

Reflecting back on the patterns observed in the data, one can see that particular groups use "breathier" /t/ variants (affricates, fricatives), while others use variants that also occur with a closed glottis (glottal stops, pre-glottalised tokens, ejectives). To start thinking further about links between these distributions, using the definitions above we categorise /t/ tokens as either "glottal", "breathy" or "canonical" (as well as "other"). While this analysis is more of an overview, it nevertheless gives a broad-brush description of the *types* of /t/-categories used. Figure 3 shows how /t/-categories used by AAE speakers (right, blue) and by MAE speakers (left, green) pattern according to these superordinate categories.

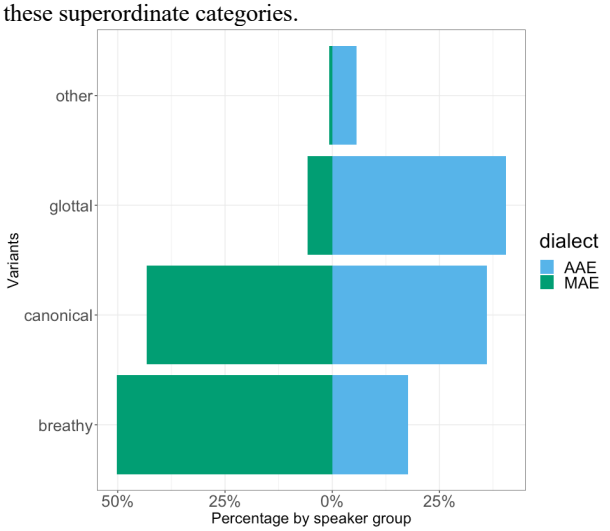


Figure 3: *back-to-back plot showing /t/ variants in broader "voice quality" categories (MAE and AAE)*

As is evident from Figure 3, AAE speakers use an overwhelmingly larger number of "glottal" /t/ types (glottal

stop, pre-glottalised, ejective), while MAE speakers use a considerably larger number of "breathy" stops (fricative, intermediate, affricate). While both groups were observed to use canonical stops, MAE speakers also use significantly more (see 3.2). Figure 4 shows the same grouping of tokens into superordinate categories, but this time according to region. Here we can see that the WN speakers use more breathy *and* more glottal /t/ variants (i.e. non-canonical), whereas the MI speakers use more canonical /t/. The differences observed across regions are not as evident as those seen for dialect.

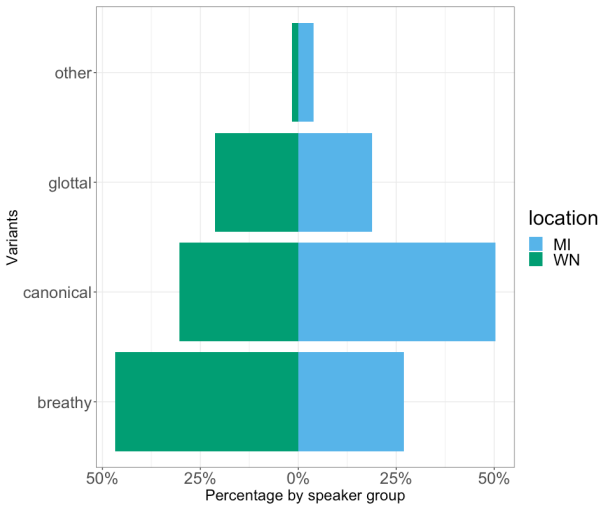


Figure 4: *back-to-back plot showing /t/ variants in broader "voice quality" categories (by region)*

4. Discussion and Conclusion

4.1. Sociophonetic patterning of /t/ in Victoria

This paper has shown both social and regional patterning of /t/ in Australian Englishes spoken in Victoria. Reflecting on our research questions, we found: 1) that regional variation is present in Australian English /t/, 2) that previously observed differences across AAE and MAE hold in a second location (MI), and 3) that both age and gender differences are evident. However, we also observed that while somewhat significant (for some tokens) gender was less strong as a sociophonetic driver of /t/ patterning.

We note that differences in the distributions of /t/ types presumably add to the overall percept of difference between the two varieties, as also noted by [15] regarding similar stop consonant variability in Croker Island English. Additionally, we saw that AAE speakers demonstrate far more variability than MAE speakers, which also aligns with findings in [15]. This is interesting given that their study included L1 and L2 speakers from a remote location in the north of Australia, whereas we focus on L1 speakers from regional (but not remote) locations in the south-east of Australia. We do not think the source of this variation is due to Aboriginal languages, but perhaps less reliance on a "standard" (an argument that applies to both L1 and L2 speakers).

Before moving on to the discussion of voice quality, some interesting observations about specific /t/ categories can be made. Here, our findings align with usage in less conservative speakers (i.e. in male speech and in the speech of younger people, and in AAE speakers more than MAE speakers). In [8] it was predicted that fricated /t/ would be more common in women's speech (similar to findings in [9]), but that is not the case here. [10] observed finer-grained social behaviour with fricative /t/, noting it relates to socioeconomic class in Perth, which also appears to be the case in Melbourne (i.e. [7]). It is also of interest that ejective and glottal stops, which are often not described as typical variants of Australian English /t/, are so prevalent for AAE speakers. However, the very fact that these stops have a potential link with laryngeal behaviour is a likely motivator for the higher frequency of these tokens in the AAE data. We will explore this idea further in future research.

Findings for some other categories are perhaps not surprising and concur with observations by other researchers. For example (pre)-glottalisation has been described by [4] as an emerging feature in MAE (that cues voicelessness in perception), and one which is used more by younger speakers as we have seen here. Studies on taps in MAE [22] note that it is an "acceptable variant", is prosodically conditioned, and tends to be used less often by "conservative" speakers, again as we have seen here (more for males, more for AAE speakers).

4.2. The link between voice quality and /t/-category

This brings us to consider broader patterns in /t/ distribution, by comparing these results with an existing analysis of voice quality for these same speakers [18]. In that study, it was found that MAE speakers exhibit breathier voice qualities than AAE speakers who in turn have voice qualities that are creakier. AAE speakers also use what may potentially be called pressed voice [18]. Also, WN speakers use significantly less modal voice, and have more breathy (and more creaky tokens) on average. Both of these points align with the findings presented in 3.5. [18] show that these AAE speakers have a voice quality where glottal constriction is used, and here we see their preferred /t/ variants also have a constricted glottis (glottal stops, pre-glottalised tokens, ejectives). MAE speakers have breathier voices, and breathier /t/ tokens (fricatives, affricates).

This very preliminary connection between voice quality and segmental findings shows that laryngeal activity appears to be conditioning certain types of supralaryngeal speaker behaviour. Results so far at least support the claim that voice quality and consonant articulation are related to some degree. This accords with observations in previous research, such as [23: 443] who questions "whether there is a connection between creaky voice and (-t) glottalization" that may be linked to a "general pattern of greater laryngeal activity". This is also taken up by [24: 18], who considers, after discussing /t/ glottalisation, "whether voice quality modulations are controlled by speakers, or are automatic consequences of other articulations", pointing out that it is likely to be both, and that these modulations can acquire social meaning (as we are likely seeing here).

For now, we can say that voice quality and supralaryngeal activity appear *not* to be totally independent, but further work is needed to better understand this. Our next step is to look closely at the dynamics of voice quality in vowels, and how it relates to each individual /t/ articulation so that we can determine the limits of this relationship. We do not think, for example, that a creaky voice quality would preclude breathy /t/ variants, or vice-versa, just that the likelihood of "glottalic" tokens is greater when voice qualities are creakier as we have seen here in a broad sense.

In summary, our paper shows a high level of sociophonetic patterning of /t/ in Australian English(es), and begins to make the important connection between one type of variability (consonant behaviour) with another (voice quality).

5. References

[1] Horvath, B.M., Variation in Australian English: The Sociolects of Sydney.Cambridge University Press, 1985.

[2] Haslerud, V.C.D., The Variable (t) in Sydney Adolescent Speech. University of Bergen, 1995.

[3] 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.

[4] Penney, J., Cox, F., and Szakay, A., "Links between production and perception of glottalisation in individual Australian English speaker/listeners", In Proc. Interspeech, 3750-3754, 2020.

[5] Ingram, J.C.L., Connected speech processes in Australian English. Australian Journal of Linguistics, 9:21-49, 1989.

[6] Tollfree, L., "Variation and change in Australian English consonants: reduction of /t/", in D. Blair and P. Collins, [Eds], English in Australia, 45-67, John Benjamins, 2001.

[7] Jones, M.J. and McDougall, K., "The acoustic character of fricated /t/ in Australian English: a comparison with /s/ and /ʃ/", Journal of the International Phonetic Association, 39(3):265-289, 2009.

[8] Loakes, D. and McDougall, K., "Individual variation in the frication of voiceless plosives: a study of Australian English speaking twins", *Australian Journal of Linguistics*, 30(2):155-181. 2010.

[9] Tait, C. and Tabain, M., "Patterns of gender variation in the speech of primary school-aged children in Australian English: the case of /p t k/", Proc. Sixteenth Australasian International Conference on Speech Science and Technology, 65-68, 2016.

[10] Docherty, G., Foulkes, P. González, S. and Mitchell, N., "Missed connections at the junction of sociolinguistics and speech processing", Topics in Cognitive Science 10:1–16, 2018.

[11] Butcher, A., "Linguistic aspects of Australian Aboriginal English", Clinical Linguistics and Phonetics, 22(8):635-642, 2008.

[12] Malcolm, I., "Australian creoles and Aboriginal English: Phonetics and phonology", in K. Burridge and B. Kortmann [Eds] Varieties of English 3: The Pacific and Australasia, 124-141, De Gruyter Mouton, 2008.

[13] Malcolm, I,. Australian Aboriginal English: Change and Continuity in an Adopted Language, De Gruyter Mouton, 2018.

[14] Dickson, G., "Aboriginal English(es)", in L. Willoughby and H. Manns [Eds] Australian English Reimagined: Structure, Features and Developments, 134-154, Routledge, 2020.

[15] 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.

[16] Loakes, D., McDougall, K., Clothier, J., Hajek, J. and Fletcher, J., "Sociophonetic variability of post-vocalic /t/ in Aboriginal and mainstream Australian English", Proc. Seventeenth Australasian International Conference on Speech Science and Technology, 5-8, 2018.

[17] Keating, P. and Esposito, C., "Linguistic voice quality", UCLA Working Papers in Phonetics, 105:85-91, 2007.

[18] Loakes, D. and Gregory, A., (2022). "Voice quality in Australian English", *JASA-EL*, 2(8):085201.

[19] Boersma, P. and Weenink, D. Praat: doing phonetics by computer [Computer program]. Version 6.2.14, http://www.praat.org/, 2022.

[20] Schiel, F., Draxler, C., and Harrington, J., "Phonemic segmentation and labelling using the MAUS technique", New tools and methods for very-large-scale phonetics research workshop, January 29–31, University of Pennsylvania, Philadelphia, PA. 2011.

[21] Winkelmann, R., Harrington, J. and Jänsch, K., "EMU-SDMS: Advanced speech database management and analysis in R", Computer Speech and Language 45:392-410, 2017.

[22] Evans, Z. and Watson, C., "Consonant Reduction in Three Dialects of English". Proceedings of the 15th ICPhS, Barcelona, 917-920, 2003.

[23] Podesva, R., "Gender and the social meaning of non-modal phonation types", Proceedings of the 37th Annual Meeting of the Berkeley Linguistics Society, ed. by C. Cathcart, I. Chen, G. Finley, S. Kang, C. Sandy and E. Stickles, 427-448, 2013.

[24] Garellek, M., "Theoretical achievements of phonetics in the 21st century: Phonetics of voice quality", Journal of Phonetics, 94:101155, 2022.

Acoustic and durational characteristics of Anindilyakwa vowels

Rosey Billington[1], John Mansfield [2], Hywel Stoakes[2]

[1]Australian National University, [2]The University of Melbourne

rosey.billington@anu.edu.au, john.mansfield@unimelb.edu.au, hstoakes@unimelb.edu.au

Abstract

Anindilyakwa, an Aboriginal language of northern Australia, has a vowel system which is unusual among Aboriginal languages, and the subject of divergent analyses. Previous research notes extensive variation in how certain vowels are produced, with observations that non-low vowels are strongly influenced by their consonant environment. There is also extensive variation in whether certain vowels are produced, leading to suggestions that these vowels are epenthetic. This study presents a first phonetic investigation of the acoustic and durational properties of Anindilyakwa vowels, with a focus on the effects of consonant place of articulation on the realisation of non-low vowels.
Index Terms: vowels, acoustics, duration, epenthesis, Australian languages, coarticulation

1. Introduction

1.1. Anindilyakwa language

The Anindilyakwa language [1] is owned and spoken by the Warnumamalya people of Groote Eylandt in northern Australia. Until recently, a prevailing view was that Anindilyakwa was a language isolate, with striking apparent differences compared to its nearest mainland neighbours. It is now viewed as a Gunwinyguan language, most closely related to Wubuy, but having undergone significant phonological and phonotactic restructuring [1]. While there have been various studies of the sound system of Anindilyakwa [2][3][4][5], there is currently no consensus on the segmental contrasts and ways these interact with specific phonological processes. This is an area of particular interest for closer investigation, drawing on new data types.

1.2. Anindilyakwa phoneme inventory

While there are varying analyses of the Anindilyakwa consonant inventory, it is broadly typical of Australian languages, with a large number of place contrasts, and more sonorants than obstruents [6]. In the inventory shown in Table 1, the 'anterior' place of articulation merges potential apical-alveolar and lamino-dental distinctions, which are not apparent in data for the present study but which have been reported elsewhere as a marginal contrast [5]. A more typologically unusual feature of the inventory is the labialised dorsal consonants /k^{w}/ and /ŋ^{w}/ (and according to some, additional labialised bilabials [4]). Some descriptions also propose a contrastive series of prenasalised stops (/mp, nt, n̪t̪, ɳʈ, ɲc, ŋk, ŋk^{w}/ [4][5]) and labial-velar double articulations (/kp, ŋm, ŋp/ [5]), but here we treat these instead as consonant sequences based on their distributions and other phonotactic analyses [7].

For the vowel system, existing proposals include just one core phonemic vowel /a/, plus marginal /e/ and phonetic [i, ə, u]

[1]ISO 639-3: aio; glottocode: anin1240

Table 1: *Anindilyakwa consonant inventory.*

	lab.	ant.	retro.	alv.-pal.	dors.	lab. dors.
stop	p	t̪ (t)	ʈ	c	k	k^{w}
nas.	m	(n̪) n	ɳ	ɲ	ŋ	ŋ^{w}
lat.		l̪ (l)	(ɭ)	ʎ		
trill		r				
appr.	w		ɻ	j		

[2]; two phonemic vowels /a, i/ [4]; and four phonemic vowels, either /a, e, i, u/ [3] or /a, ɛ, i, ə/ [5]. Regardless of the phonemic analysis, there is broad agreement that vowel phones in Anindilyakwa include two 'low' vowels [ɛ, a], and three 'non-low' vowels [i, ə, u]. In part, the differing analyses are due to observations that the production of the 'non-low' vowels, in terms of frontness and rounding, is conditioned to at least some extent by the place of articulation of neighbouring consonants. As observed by Heath [2] and subsequent analysts, non-low vowels are generally realised as [u] when adjacent to a labial or labialised dorsal consonant, and [i] when adjacent to an alveo-palatal consonant. Elsewhere, they are generally realised as [ə]. At the same time, various lexical exceptions have been reported, especially in instances where [i] has no conditioning palatal, such as [aɻimpa] 'stingray sp.' [3] and [mipina] 'same' [5], and some analysts report fluctuation between different qualities in the same environments and words [2]. There is also some evidence for conditioning in the low vowels, where [ɛ] frequently appears adjacent to a palatal and is viewed by some as an allophone of [a] [4], but elsewhere found to be contrastive [5]. The apparent predictability but also variation in non-low vowel quality suggests a need for closer examinination of the relationship between non-low vowel phones [i∼ə∼u] and consonant context. Proposals for the Anindilyakwa vowel system differ from the typologically more common 'triangular' 3–5 vowel systems found in many Australian languages, with /i, (e), a, (o), u/ (and often length contrasts) [8]. However, some similar analytical challenges posed by interactions between vowel quality and consonant environment can be found in Arandic languages such as Kaytetye and Central Arrernte, with arguably 'vertical' small vowel systems in which height is the primary parameter of contrast (e.g. [9], [10]).

1.3. Predictability of vowel occurrence

Previous researchers also note that beyond vowel quality, there is variation in whether or not non-low vowels are produced in certain contexts, and Heath [2] argues that the phones [i, ə, u] only occur as 'brief' interconsonantal epenthetic vowels, and are largely predictable in where they occur. Vowel epenthesis, which broadly relates to the surface insertion of vocalic segments and manifests in many ways crosslinguistically

[11], is uncommon among Australian languages.[2] A recent information-theoretic analysis investigating the predictability of vowel occurrence across different consonant environments in Anindilyakwa, based on both orthographic representations in a wordlist [12] and on the corpus of segmented production data used in the current study, finds that there is indeed a high level of predictability in the occurrence of 'non-low' vowels compared to 'low' vowels depending on the manner and place of adjacent consonants, for example that putative epenthesis is rare in sequences of dorsal and labial consonants and in homorganic nasal and stop sequences, but extremely common in other environments, such as in consonant sequences of equal or increasing sonority. However, patterns in the presence/absence of non-low vowels are more variable in the production data than in the wordlist data, in that the same lexeme may be produced in different ways, and in that non-low vowels may be omitted in contexts where they appear in the orthographic wordlist [7].

2. Research aims

The present study builds on the investigation of the presence compared to absence of non-low vowels in Anindilyakwa speech production data, and examines the phonetic characteristics of these vowels. In particular, we use formant measurements to investigate the claim that non-low vowel quality is largely influenced by the place-of-articulation of neighbouring consonants. Secondly, we use durational measurements to test the claim that non-low vowels are 'brief' in comparison to low vowels. The aim is to develop the understanding of how production patterns for Anindilyakwa vowels accord with impressionistic descriptions of vowel realisation in the language, and lay the groundwork for targeted research on coarticulation and natural speech processes.

3. Method

3.1. Participants

We present data collected on Groote Eylandt with seven Anindilyakwa speakers, five women and two men. The speakers' ages range from approximately 25 to 80 years old. They all speak Anindilyakwa as their main daily language, and all are multilingual in Kriol, English and other regional languages (especially Wubuy and Yolngu Matha).

3.2. Materials and procedures

Due to the varying phonological analyses (and orthographic conventions) for Anindilyakwa as well as the reported variation in vowel quality and presence, existing materials are highly inconsistent in the representation of lexemes, presenting a challenge for stimuli design. Therefore, data collection for this exploratory phonetic study (and associated study of vowel presence/absence [7]) focuses on eliciting naturalistic utterances likely to represent a broad range of segmental combinations. A set of target nouns was prepared, and elicited using picture prompts and in some cases spoken English prompts. Speakers were audio-recorded producing the target nouns in utterance frames they found meaningful, in accordance with community preferences for the conduct of this study. The same prompts were used for each speaker, meaning there is some comparability across the target nouns, but as the speakers were not required to use specific sentence frames, there is substantial diversity in how they chose to respond to the prompts. The number of utterances collected for each speaker ranged from 53–111, with the exception of one female speaker who produced 24 utterances.

[2]The prevalence of word-final /a/ in Anindilyakwa is also argued to be related to a separate epenthesis process [5].

3.3. Data processing and analysis

Utterances were orthographically transcribed and used to create an EMU-SDMS hierarchical database [13] [14] [15], following conversion to IPA and automated phone segmentation via the Australian Aboriginal Language model [16] in WebMAUS [17]. Segmentation was manually checked and corrected across the data. The database contains a total of 473 utterances and 5593 vowel tokens (see Table 2). Low vowel tokens (N=3674) were labelled [ɐ] and [ɛ] by the authors according to perceived vowel quality, recalling that previous phonological analyses agree that there is a contrastive open central vowel, and most likely a marginal mid front vowel. Non-low vowel tokens (N=1919) were labelled [ɪ, ə, ʊ] according to perceived vowel quality, recalling that previous analyses, while varied, posit these differing vowel qualities as largely arising from consonantal effects.

Table 2: *Number of vowel tokens in dataset.*

height	quality	# tokens
'low'	[ɐ]	3165
	[ɛ]	509
'non-low'	[ɪ]	471
	[ə]	843
	[ʊ]	605

4. Results

4.1. Formant frequency

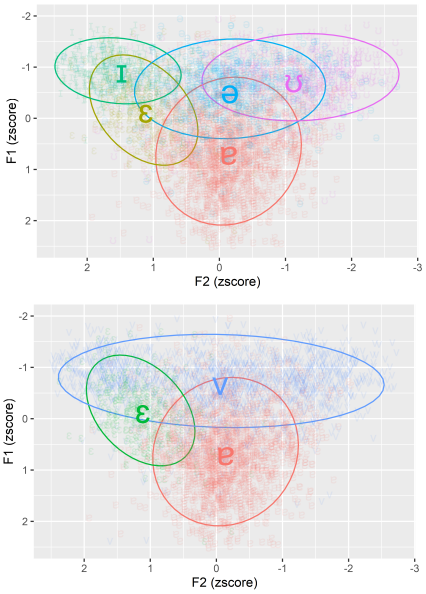


Figure 1: *F1 and F2 midpoints (Lobanov-normalised), by labelled vowel quality (top) and with non-low vowels grouped together (bottom).*

Normalised first and second frequency measures, based on vowel midpoints, are shown in Figure 1 (excluding all tokens with zero values for F1, F2 or F3). In the top panel, vowels are plotted according to labelled vowel quality, in line with

the five vowel phones which have been reported by all previous researchers, regardless of their phonemic analysis. While a lot of variation is apparent, distributions broadly correspond to vowel realisations which are open central [ɐ], mid front [ɛ], close front [ɪ], mid (or close) [ə/i], and close back [ʊ]. With all non-low vowels grouped together as 'v' (bottom panel), in line with the most minimalist analyses of all non-low vowels as either epenthetic [2] or allophones of a single phoneme [4], we can see that their distributions are concentrated in the upper part of the vowel space, and that variation within this grouping, as well as across the three groupings in the top panel, is largely in the front-back dimension. For [ɐ], mean F1 is 726Hz (s.d. 199) and mean F2 is 1697Hz (s.d. 235) for the female speakers, and mean F1 is 625Hz (s.d. 165) and mean F2 is 1435Hz (s.d. 259) for the male speakers. For [ɛ], mean F1 is 540Hz (s.d. 95) and mean F2 is 2193Hz (s.d. 167) for the female speakers, and mean F1 is 516Hz (s.d. 140) and mean F2 is 1892Hz (s.d. 230) for the male speakers. For all non-low vowel phones grouped together, mean F1 is 421Hz (s.d. 85), and mean F2 is 1717Hz, with a large standard deviation of 478, for the female speakers, and mean F1 is 442Hz (s.d. 143) and mean F2 is 1504Hz, with a similarly large standard deviation of 498, for the male speakers.

Table 3: *Mean (and s.d.) F1 and F2 at midpoints for non-low vowel tokens produced by female (f) and male (m) speakers, by consonant context (*=only one token for (m) in this context).*

context	N	F1 (f)	F2 (f)	F1 (m)	F2 (m)
LBD_LBD	3	434 (30)	913 (293)	580 (-)	838 (-) *
LAB_LBD	22	411 (83)	907 (231)	464 (92)	1408 (785)
DOR_LBD	17	429 (69)	865 (179)	463 (82)	689 (124)
ANT_LBD	138	414 (58)	1264 (309)	483 (145)	1103 (445)
PAL_LBD	65	399 (45)	1627 (257)	363 (61)	1392 (313)
LBD_LAB	14	405 (56)	948 (205)	473 (-)	2208 (-) *
LAB_LAB	5	434 (23)	1087 (11)	378 (130)	1374 (407)
DOR_LAB	54	420 (98)	1320 (305)	522 (248)	1333 (485)
ANT_LAB	150	402 (64)	1650 (305)	476 (147)	1381 (410)
PAL_LAB	59	381 (59)	2056 (329)	347 (72)	1774 (370)
LAB_DOR	20	407 (128)	1139 (360)	354 (-)	2293 (-) *
DOR_DOR	3	487 (177)	1906 (132)	-	-
ANT_DOR	238	441 (97)	1832 (358)	487 (136)	1674 (444)
PAL_DOR	91	389 (54)	2411 (250)	404 (84)	2120 (224)
LBD_ANT	127	448 (93)	1178 (226)	451 (162)	1057 (342)
LAB_ANT	161	423 (75)	1447 (299)	403 (94)	1364 (434)
DOR_ANT	166	473 (101)	1738 (245)	577 (195)	1510 (373)
ANT_ANT	142	435 (100)	1835 (190)	435 (89)	1355 (294)
PAL_ANT	98	429 (65)	2204 (259)	384 (92)	1928 (275)
LBD_PAL	53	363 (58)	1624 (342)	336 (90)	1268 (256)
LAB_PAL	66	374 (38)	2189 (321)	352 (68)	1837 (380)
DOR_PAL	13	394 (52)	2292 (553)	410 (85)	2136 (97)
ANT_PAL	52	386 (53)	2202 (293)	361 (51)	2051 (148)
PAL_PAL	41	352 (67)	2406 (283)	317 (58)	2219 (110)

The realisation of non-low vowels in the front-back dimension is examined in more detail in Figure 2. All non-low vowel tokens are plotted according to consonant context, here the place of articulation of the preceding and following consonant.[3] As can be seen in Table 3, some homorganic consonant contexts correspond to very few tokens; these are shown here for completeness, but acoustic measures for these contexts are naturally not very meaningful at this stage. The differing token numbers relate to the varying presence/absence of non-low vowels depending on consonant manner and place, as discussed in Section 1.3. As can be seen, F1 and F2 patterns for non-low vowels are highly gradient; different consonantal contexts do not produce discrete groups, but rather a continuum of realisations ranging from higher F2 values in palatal environments towards the left, and lower F2 values in labio-dorsal environments towards the right (see Table 3). Statistical tests using linear mixed-effects models via `lme4` [18], with random intercepts for speaker and word, indicate that there is a significant effect of consonant environment on F2 for the non-low vowels ($p<0.001$), whether the fixed effect is the C1_C2 context, or just the following C, or just the preceding C. In Tukey-adjusted post-hoc pairwise comparisons, there are significant differences for the majority of place of articulation comparison environments. For F1, the effect of consonant environment is also significant ($p<0.001$), whether treated as C1_C2 context, following C, or preceding C, and post-hoc tests indicate significant differences particularly when the following consonant is palatal, and when the preceding consonant is palatal or labial/labialised dorsal compared to anterior or dorsal.

Inspection of lexemes with non-low vowels shows that while the overall effects of consonant context are more or less as expected based on previous descriptions, for example with [ɪ] occurring after a palatal glide and [ʊ] after a bilabial stop in [jɪpʊɻɐt̪ɐ] 'wallaby', there are also exceptions. Some lexical exceptions appear to be acoustically variable: for example, tokens of the word 'good' are variably produced as [ɛnɪŋɐpɐ], with [ɪ] in a non-palatal context, or as [ɛnəŋɐpɐ], with [ə] between the palatal and dorsal nasals. But there also appear to be more robust lexical exceptions: for the word 'back', tokens are quite consistently produced as [məɻɪrpɐ], with close front [ɪ] as the non-low vowel between the trill and retroflex approximant, suggesting that an unconditioned /ɪ/ vowel is part of the lexical representation of this word.

4.2. Duration

Normalised duration values can be seen in Figure 3, excluding pre-pausal vowels (which may exhibit final lengthening). Distributions indicate that the non-low vowels are typically much shorter than the low vowels. The mean duration for the low vowel [ɐ] is 103ms (s.d. 45), and for [ɛ] the mean duration is 96ms (s.d. 35). For the non-low vowel qualities grouped together, as shown in the right panel, the mean duration is 52 ms (s.d. 24), meaning that the low vowels are on average approximately twice as long as the non-low vowels. Means for the non-low vowels are similar if separated according to labelled vowel quality, as shown in the left panel; for [ɪ], 61ms (s.d. 28), for [ə], 48ms (s.d. 21), for [ʊ], 53ms (s.d. 21). Statistical tests using linear mixed-effects models with random intercepts for speaker and word indicate that there is a significant effect of vowel quality on duration (whether the fixed effect of vowel quality is non-low vs. low vowels or the five labelled vowel qualities), and Tukey-adjusted post-hoc pairwise comparisons confirm the differences between [ɐ, ɛ] and [ɪ, ə, ʊ] ($p<0.001$).

5. Discussion and conclusions

Our findings for first and second formant frequency accord with impressionistic observations that vowel production in Anindilyakwa broadly corresponds to vowel qualities [ɐ, ɛ, ɪ, ə, ʊ] [2][3][4][5]. At the same time, substantial variation is apparent, with highly gradient patterns for non-low vowels depending on the consonantal context. While substantial variation in vowel realisation, particularly due to consonantal effects,

[3] Our data included only small numbers of retroflex consonants, which have been grouped as anterior here.

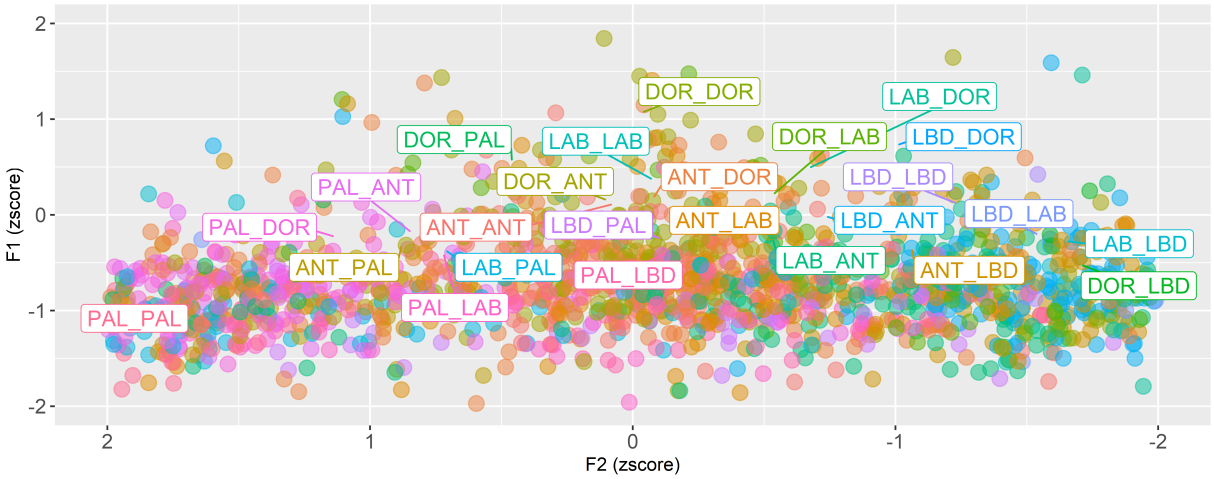


Figure 2: *F1 and F2 midpoints (Lobanov-normalised) for non-low vowel tokens, by consonant context.* LBD=*labialised dorsal,* LAB=*labial,* DOR=*dorsal,* ANT=*anterior*, PAL=palatal.

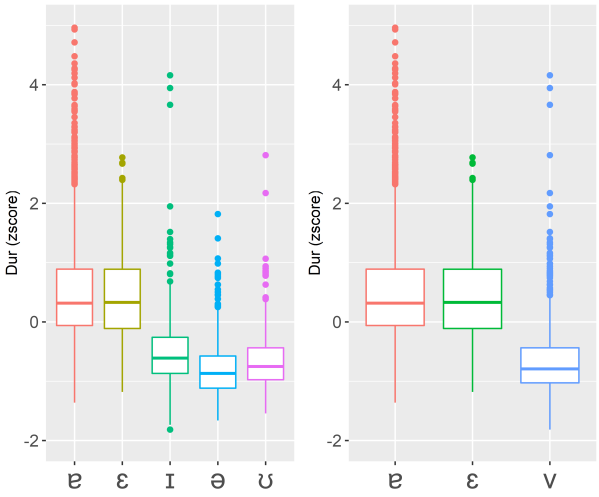


Figure 3: *Duration (Lobanov-normalised), by labelled vowel quality (left) and with non-low vowels grouped together (right).*

is common for Australian languages with small vowel systems [8], in Anindilyakwa this has posed particular analytical challenges, and the arguably epenthetic status of the non-low vowel phones is a notable typological difference. However, although Anindilyakwa has undergone dramatic phonological restructuring [1], the resulting system may still exhibit a common pattern proposed for Australian languages, the 'place of articulation imperative' [6], whereby various phonetic and phonological patterns are optimised for maintaining perceptual cues to consonant place distinctions; the non-low vocalic segments in Anindilyakwa may well have such a role (regardless of whether these vowels are interpreted as epenthetic [2] or allophones of one phoneme [4]). Some lexical items in this dataset showed [ɪ] occurring in environments other than adjacent to alveo-palatal consonants, as also reported elsewhere [3] [5], aligning with impressions of a marginal contrast.

The duration results for the low vs. non-low vowels reinforce impressions that these two sets of vowel phones pattern differently in the Anindilyakwa sound system. While open vowels, crosslinguistically, tend to be somewhat longer than close vowels, here the Anindilyakwa low vowels are twice as long as the non-low vowels, a larger difference than might be expected on the basis of biomechanical factors rather than when duration is a cue to vowel distinctions [19] [20]. These findings also match impressions reported by [2] that the non-low vowels are 'very brief' and the low vowels are 'considerably longer'; in fact, Heath estimates they are 'normally at least twice the duration'. Van Egmond [5] similarly notes that while there is no contrastive length, the low vowels are characteristically longer than the non-low vowels. Shorter durations are not uncommon for epenthetic vocalic segments [11]. The difference between the low and non-low vowels here is also larger than the reported duration difference between contrastive long and short open vowels in closely-related Wubuy, where /aː/ is 1.28 times longer than /a/ [21].

A matter that remains to be investigated is the role of prosody, and how word-level prosodic patterns interact with vowel quality and duration as well as vowel presence/absence. Previous discussions of stress in Anindilyakwa claim a predominant pattern of penultimate primary stress [4] [5], but also indicate that stress may be quantity-sensitive, in that the low vowels, while not analysed as contrastively long, attract stress due to their longer durations, and closed syllables likewise attract stress [5]. Where these factors compete, stress may vary. However, [5] notes that these are preliminary analyses based on isolated word forms. While word prosody has not yet been closely investigated for the present data, exploratory examinations indicate that pitch peaks are typically located towards the left edges of words, sometimes on syllables containing non-low vowels, suggesting that the analysis of word prosody needs to be revisited. Given that non-low vowels which arguably function as epenthetic are suggested to historically derive from centralisation of full vowels *i, *u and *a in coarticulatory and unstressed environments [1], further investigation of prosodic patterns may shed light on both synchronic and diachronic processes.

This exploratory investigation of Anindilyakwa vowel quality, building on analyses of non-low vowel presence vs. absence, bolsters the evidence that both the occurrence and the quality of non-low vowels is to a large extent predictable, based on the surrounding consonant environment. Future work will benefit both from analyses of controlled production data for vowel phones, in known epenthesis environments and with different combinations of consonant place of articulation, as well as analyses of spontaneous speech data, in order to better understand natural speech processes and individual speaker behaviour in Australian languages with small vowel systems.

6. Acknowledgements

We wish to thank all the Anindilyakwa speakers who taught us about their language, and especially the seven who spoke sentences for this study: Ramesh Lalara, Kathleen Mamarika, Katelynn Bara, Edith Mamarika, Judy Lalara, Coleen Mamarika and Joel Marawili. We also thank staff at the Groote Eylandt Language Centre – James Bednall, Brighde Collins and Carolyn Fletcher – who provided generous support and hospitality to the second author in his field visits. We gratefully acknowledge funding support from the University of Melbourne Research Unit for Indigenous Language, and the Australian Research Council grant DE180100872.

7. References

[1] van Egmond, M.-E. and Baker, B. "The genetic position of Anindilyakwa", Australian Journal of Linguistics, 40(4):492–527, 2020.

[2] Heath, J. "Draft grammatical sketch of Anindhilyagwa".

[3] Stokes, J. "Anindilyakwa phonology from phoneme to syllable". In Waters, B., editor, *Work Papers of SIL-AAIB, Series A Volume 5*, pages 139–181. SIL, Darwin, 1981.

[4] Leeding, V. J. Anindilyakwa phonology and morphology. PhD thesis, University of Sydney, 1989.

[5] van Egmond, M.-E. Enindhilyakwa phonology, morphosyntax and genetic position. PhD thesis, University of Sydney, 2012.

[6] Butcher, A. "Australian Aboriginal languages: Consonant-salient phonologies and the 'place of articulation imperative"'. In Harrington, J. M. and Tabain, M., editors, *Speech production: Models, phonetic processes and techniques*, pages 187–210. Psychology Press, New York, 2006.

[7] Mansfield, J., Billington, R., and Stoakes, H. "Vowel predictability and omission in Anindilyakwa". PsyArXiv, pre-print.

[8] Fletcher, J. and Butcher, A. "Sound patterns of Australian languages". In Koch, H. and Nordlinger, R., editors, *The languages and linguistics of Australia: A comprehensive guide*, pages 91–138. De Gruyter Mouton, Berlin, 2014.

[9] Breen, G. "The wonders of Arandic phonology". In Simpson, J., Nash, D., Laughren, M., and Alpher, B., editors, *Forty years on: Ken Hale and Australian languages*, pages 45–69. Pacific linguistics, Canberra, 2001.

[10] Tabain, M. and Breen, G. "Central vowels in Central Arrernte: A spectrographic study of a small vowel system", Journal of Phonetics, 39(1):68–84, 2011.

[11] Hall, N. "Vowel epenthesis". In Oostendorp, M., van, Ewen, C. J., Hume, E., and Rice, K., editors, *The Blackwell companion to phonology*, volume 3, pages 1576–1596. Wiley, 2011.

[12] Waddy, J. A. "Draft Anindilyakwa dictionary". 1989.

[13] Winkelmann, R., Harrington, J., and Jänsch, K. "EMU-SDMS: Advanced speech database management and analysis in R", Computer Speech & Language, 45:392–410, 2017.

[14] Winkelmann, R., Jänsch, K., Cassidy, S., and Harrington, J. emuR: Main package of the EMU Speech Database Management System, 2021. R package, Version 2.3.0.

[15] R Core Team. R: A language and environment for statistical computing, 2022. Version 4.2.0.

[16] Stoakes, H. and Schiel, F. A Pan-Australian model for MAUS, 2017.

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

[18] Bates, D., Mächler, M., Bolker, B. M., and Walker, S. C. "Fitting linear mixed-effects models using lme4", Journal of Statistical Software, 67(1):1–48, 2015.

[19] Lindblom, B. "Vowel duration and a model of lip mandible coordination", Speech Transmission Laboratory Quarterly Progress and Status Report, 8(4):1–29, 1967.

[20] Solé, M.-J. and Ohala, J. "What is and what is not under the control of the speaker: Intrinsic vowel duration". In Fougeron, C., Kühnert, B., D'Imperio, M., and Vallée, N., editors, *Papers in Laboratory Phonology 10*, pages 607–655. De Gruyter Mouton, Berlin, 2010.

[21] Bundgaard-Nielsen, R. and Baker, B. "The vowel inventory of Roper Kriol". In The Scottish Consortium for ICPhS 2015, editor, *Proceedings of the 18th International Congress of Phonetic Sciences*, pages 1–4 (paper 0885). University of Glasgow, Glasgow, 2015.

The Vowel Inventory of the Kufo Language

Shubo Li

The Australian National University, ARC Centre of Excellence for the Dynamics of Language

shubo.li@anu.edu.au

Abstract

The linguistic diversity of East Africa is well-known, but understandings of the languages in some regions, such as the Nuba Mountains of Sudan, remain limited. Based on phonological evidence and acoustic phonetic data, this paper presents a preliminary description of the vowel inventory of Kufo, a Kadu language. Vowel contrasts based on both quality (including an Advanced Tongue Root distinction) and length have been previously proposed for Kufo, and the evidence for these is examined via measurements of first and second formant frequencies and duration.

Index Terms: Kufo, vowels, ATR, F1, F2, duration

1. Introduction

Kufo (also known as Kufa or Kufa-Lima) is a variety of the Kanga language [1] which is traditionally spoken in the Nuba Mountains in South Kordofan, Sudan. The estimated number of Kanga speakers is 8,000 [2]. Kanga is classified as part of the Kadu language family. No higher-level genetic affiliations have been established for this family, though some early studies viewed it as part of the Niger-Congo phylum, and more recent work suggests it has more in common with languages of the disputed Nilo-Saharan phylum [3] [4] [5] [6]. The current study is part of a wider documentation project taking place with one Kufo speaker living in Australia, and presents results based on phonological observations and an acoustic analysis investigating the Kufo vowel inventory.

1.1. Existing proposals for the Kufo vowel system

Current descriptive research on Kufo and related varieties is extremely limited, largely comprising some wordlist materials and preliminary phonological observations. The vowel inventory has not yet been investigated with comprehensive phonological data, nor any supporting phonetic evidence. Existing phonological observations include proposals that the Kufo vowel system may include a contrast based on an 'Advanced Tongue Root' (ATR) distinction, and that at least some Kufo vowels contrast in length. However, there are different views on the number of monophthongs in Kufo, and the nature of the ATR contrast remains unclear (and there is currently no clear indication of any patterns of vowel harmony drawing on this feature, as is often found in other languages). Schadeberg [4] examines nine Kadu languages and proposes a basic seven vowel system (/i, ɪ, ɛ, a, ɔ, ʊ, u/) with a length contrast for all vowels. In comparison, in an overview of eight Kadu languages, Hall & Hall [5] suggest that the Kufo vowel inventory consists of nine vowels including four [+/-ATR] pairs (/i, ɪ/, /e/, ɛ/, /o, ɔ/, and /u, ʊ/), as well as an open central vowel, /a/. Hall & Hall suggest that /a/ may contrast with /ə/ in some Kadu languages, but that this has not been established for Kufo. Hall & Hall also suggest that vowel length is contrastive in the Kadu languages. Based on an annotated wordlist of the Kufo language, Blench & Mongash [8] suggest that Kufo has ten phonemic vowels which are five [+/-ATR] pairs (/i, ɪ/, /e, ɛ/, /a, ə/, /o, ɔ/, and /u, ʊ/). They note that /a/ and /ə/ are clearly phonemes (though represented with the same grapheme). The length contrast is not discussed in their work but is displayed in their transcription of the wordlist. Lastly, recent work by Evans et al. [9] suggests that Kufo has an eleven-vowel inventory which includes five [+/-ATR] pairs (/i, ɪ/, /e, ɛ/, /æ, ɐ/, /o, ɔ/, and /u, ʊ/), as well as one long vowel /aː/. More recent work leading up to the present study suggests a vowel inventory with nine contrastive vowel qualities.

1.2. Nilo-Saharan vowel systems

While Kufo and the Kadu family have not yet been conclusively linked to other languages of the disputed Nilo-Saharan phylum, the proposed features of the vowel system are widely attested among other languages described as Nilo-Saharan, and of East Africa more generally.

1.2.1. Advanced Tongue Root

The Advanced Tongue Root feature, which is stated to be exhibited in Kufo in various previous studies, is widely attested among African languages. An ATR contrast is a binary distinction between vowels of similar height, backness, and rounding, held to correlate with different pharyngeal and laryngeal articulatory settings [10]. Existing phonetic investigations of ATR contrasts remain limited, and mainly focus on Niger-Congo languages of West Africa, with few phonetic studies of East African languages [11] [12]. The phonetic investigations suggest that the most crosslinguistically reliable acoustic correlate is that vowels classed as [+ATR] tend to have a lower first formant frequency (F1) than their [-ATR] counterparts. Other evidence, such as differences in second formant frequency (F2), duration, and voice quality, are less consistent [10]. For languages with an ATR contrast, a 9-vowel system is most common, where the close and mid vowels contrast but only one open vowel exists [13]. In the current project, auditory impressions are that an ATR-type contrast may be present for close and mid vowels, but there is as yet no auditory of phonological evidence for /ə/ as a contrastive [+ATR] counterpart to /a/, though phonetic [ə] does occur in restricted morphophonological contexts, namely the first syllable of conjugated verbs with three or more syllables.

1.2.2. Vowel length

Vowel length contrasts are typologically uncommon among African languages. While vowel length contrasts have been proposed for some East African languages, the phonetic evidence for these contrasts has not been examined for many languages

outside the Western Nilotic family [14] [15]. For Kufo, vowel length has been briefly mentioned in previous studies, but there are differing views on the status of length within the vowel system. In the current project, auditory impressions are that length contrasts may be present for all vowel qualities. Evidence for the contrast between long and short vowels is most apparent in the first syllable of disyllabic words; long vowels rarely appear in other environments in the currently available data.

2. Research aims

This paper seeks to further develop the description of the Kufo vowel system based on phonological evidence and analyses of the acoustic and durational characteristics of vowels. The following questions will be addressed in this paper: How many contrastive vowel qualities are there in Kufo? Is there acoustic evidence for an ATR-type contrast, based on measures of first and second formant frequency? Is there durational evidence that Kufo vowels contrast in length?

3. Method

3.1. Participant

The data in this database is collected with one male diasporic native Kufo speaker, Haroun Kafi, who was born in the 1960s and currently resides in Australia. Besides Kufo, Haroun also speaks Sudanese Arabic and English.

3.2. Materials and procedures

3.2.1. Materials

Based on the descriptive phonological data leading up to the current study, auditory impressions are that there are nine contrastive vowel qualities in Kufo, including a close front vowel /i/, a near-close front vowel /ɪ/, a close-mid front vowel /e/, an open-mid front vowel /ɛ/, an open central vowel /a/, an open-mid back vowel /ɔ/, a close-mid back vowel /o/, a near-close back vowel /ʊ/, and a close back vowel /u/. Phonemic evidence suggests a length contrast for all nine vowel qualities. The hypothesised Kufo vowel inventory is presented below.

/i/	/ɪ/	/e/	/ɛ/	/a/	[ə]	/ɔ/	/o/	/ʊ/	/u/
/iː/	/ɪː/	/eː/	/ɛː/	/aː/	-	/ɔː/	/oː/	/ʊː/	/uː/

Based on this hypothesised vowel inventory, a wordlist was developed, consisting of 66 disyllabic words with a CVCV structure. As is discussed, phonetic [ə] occurs in restricted morphophonological contexts only with no evidence for contrast, and is thus not included in the database. The wordlist was compiled based on notes from the 2021 Field Methods course at the Australian National University [16], follow-up elicitation sessions in early 2022 [17], and the unpublished wordlist by Blench and Mongash [8]. The wordlist was finalised in consultation with the speaker in early 2022, with careful consideration to include all hypothesised vowel qualities. A balance across vowel qualities was aimed for as much as possible within the limitations of the small amount of available lexical data for the language. The wordlist aimed to include words exhibiting the same vowel quality for vowels in both syllables. The CVCV words also included consonants of a range of places and manners of articulation, and had a range of tonal patterns. The items in the wordlist were arranged in a pseudo-random order to ensure that words with

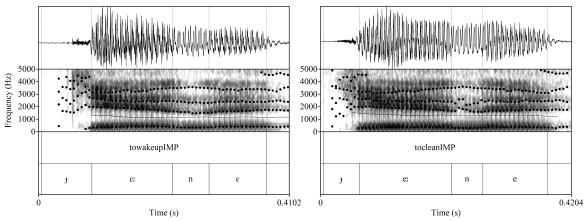


Figure 1: A vowel quality minimal pair: *ɟɛːnɛ* and *ɟeːne*.

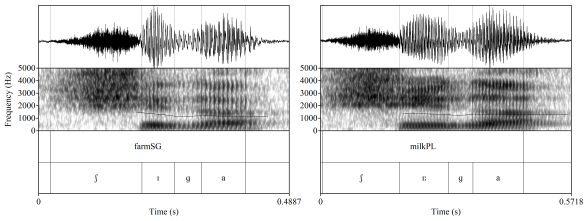


Figure 2: A vowel length minimal pair: *ʃɪga* and *ʃɪːga*.

vowels of different qualities are evenly distributed throughout the wordlist. Examples of target words include minimal pairs exhibiting a proposed vowel quality contrast, e.g., *ɟɛːnɛ* 'Wake up!' and *ɟeːne* 'Clean!', and minimal pairs exhibiting a proposed vowel length contrast, e.g., *ʃɪga* 'farm' and *ʃɪːga* 'milk'. The minimal pairs are presented in Figures 1 and 2.

3.2.2. Data collection

Recording sessions were held in a quiet room at the speaker's home, and the recordings were collected with a Zoom H6 portable audio recorder and a Rode NT3 microphone, at the archival sampling rate of 96kHz and 24-bit depth. The wordlist was arranged into Microsoft PowerPoint slides where each word was presented with its English translation and an accompanying picture. The Kufo orthography was not included in the presentation in order to avoid any influence of orthographic representations on the speaker's production. In addition, the Kufo orthography is not yet fully developed, especially regarding the representation of the Advanced Tongue Root feature, which is related to one of the aims of this study. All words in the wordlist were elicited five times consecutively in the frame *aʔa nɪkːi ... ɓɪtɛnɪ* 'I say … today', in order to control for prosodic effects and provide sufficient tokens for phonetic explorations.

3.3. Data processing and analysis

3.3.1. Data processing and annotation

The recorded data (.wav files) was down-sampled to a rate of 44.1kHz and 16 bit-depth for acoustic analysis, and then segmented and annotated in Praat [18] as .TextGrid files. Data was annotated at word and phoneme levels. The word tier contains the English translation of the Kufo word, including the number of the nouns (SG for singular and PL for plural) and the mood of the verbs (INF for infinitive and IMP for imperative). The phonemic tier contains the segmental annotation of each word using SAMPA [19].

Given that every word in this database has a CVCV structure, all vowels occur in either word-medial or word-final position. Word-medial vowels were segmented based on the onset

of periodicity after the preceding consonant, and the offset of periodicity before the following consonant. Word-final vowels were segmented based on the onset of periodicity after the preceding consonant, and either the offset of periodicity before the following bilabial implosive in the frame sentence, or, in cases where the target word is followed by a pause, based on the offset of periodicity and continuous formant structure [20].

3.3.2. Database building

The .wav files and .TextGrid files were used to create a hierarchical database with the EMU Speech Database Management System [21]. Measures of first and second formant frequencies at vowel midpoints, and of vowel duration, were extracted and analysed for vowels in the target words, using the emuR package [22] in R [23] via RStudio [24]. In total, this phonetic database consists of 333 occurrences of 66 target words. A total of 653 vowel tokens were analysed, including 328 vowels in the first syllable and 325 vowels in the second syllable. The number of tokens of each vowel is listed below in Table 1 according to the vowel quality (vq) and the syllable in which it occurs (s1 or s2 for first or second syllable). Short vowel tokens (n=532) appear in both s1 and s2, whereas long vowel tokens (n=121) appear in s1 only. There were thirteen 0 values for the second formant frequency measurement due to tracking errors, and these rows of data were removed from the dataset.

Table 1: *Number of tokens in the dataset, by vowel quality.*

vq	short		long	total	vq	short		long	total
	s1	s2	s1			s1	s2	s1	
i	35	51	20	106	ɪ	35	40	10	85
e	0	20	15	35	ɛ	0	34	29	63
o	25	60	20	105	ɔ	23	38	15	76
u	30	20	0	50	ʊ	34	25	0	59
					a	25	37	12	74

4. Results

4.1. First and second formant frequencies

The results for first and second formant frequency measurements based on vowel midpoints are displayed in Table 2 and Figure 3 according to vowel quality, with short and long vowels combined.

Table 2: *Mean and standard deviation for F1 and F2 (Hz) based on vowel midpoints.*

vq	F1	sd	F2	sd	vq	F1	sd	F2	sd
i	314	24	2214	135	ɪ	356	24	2164	150
e	399	27	1951	66	ɛ	473	51	1804	85
o	420	40	946	88	ɔ	480	52	920	126
u	340	19	1046	292	ʊ	368	24	965	204
					a	570	74	1452	230

For the close and near-close front vowel qualities, the mean F1 for /i/ and /ɪ/ are respectively 314 Hz and 356 Hz. For the close-mid and mid front vowel qualities, the mean F1 for /e/ and /ɛ/ are 399 Hz and 473 Hz. For the close and near-close back vowel qualities, the mean F1 for /u/ and /ʊ/ are respectively 340 Hz and 368 Hz. For the close-mid and open-mid back vowel qualities, the mean F1 for /o/ and /ɔ/ are 420 Hz and 480

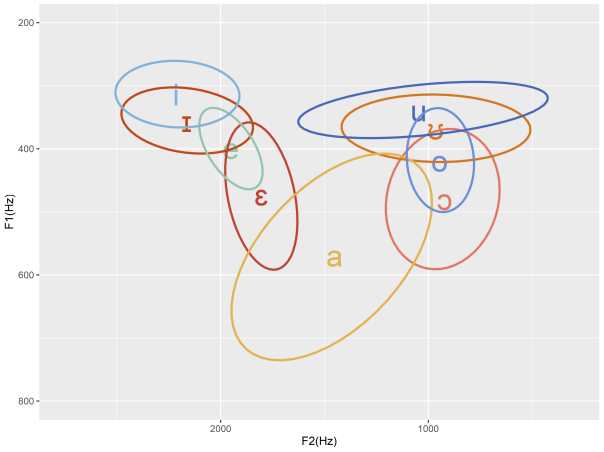


Figure 3: *First and second formant frequencies.*

Hz. The open central vowel /a/, which is not currently hypothesised to have a [+ATR] counterpart, is distinct from the other vowel qualities and has a mean F1 of 570 Hz. There are no major F2 differences between pairs of similar vowels, though there are indications that some [+ATR] vowels may tend towards a more front quality than their [-ATR] counterparts, e.g., the mean F2 of /e/ is 1951Hz, which is higher than the mean F2 of /ɛ/, 1804Hz.

The acoustic results provide some supporting evidence for an analysis of 9 contrastive vowel qualities in Kufo. While there is some overlap in the distributions for the vowel categories, the phonetic patterns accord with the current phonological hypothesis, and are also similar to vowel spaces for other 9-vowel systems with an ATR contrast.

The mean F1 values for close and mid vowels also provide some supporting evidence for an ATR-type contrast, where the closer vowel in each pair, which is potentially categorised as the [+ATR] vowel, has a lower first formant frequency than its more open counterpart, which is potentially categorised as the [-ATR] vowel. The contrasts suggest that there are likely four [+/-ATR] vowel quality pairs in Kufo: /i, ɪ/, /e, ɛ/, /u, ʊ/, and /o, ɔ/.

4.2. Vowel duration

The results for vowel duration measurements are displayed in Table 3 and Figure 4. Only the durations of vowels in the first syllables of target words are shown here, given that, as noted, this is the environment where there is the most evidence for the length contrast.

Table 3: *Mean and standard deviation for duration (ms) for vowels in initial syllables.*

v	dur	sd	v	dur	sd	v	dur	sd	v	dur	sd
i	67	8	iː	134	14	ɪ	69	15	ɪː	130	16
e	-	-	eː	146	25	ɛ	-	-	ɛː	158	15
o	82	17	oː	145	13	ɔ	81	17	ɔː	151	17
u	67	14	uː	-	-	ʊ	69	12	ʊː	-	-
						a	79	10	aː	157	27

For the [+ATR] close front vowels, the mean duration of the short vowel /i/ and the long vowel /iː/ are respectively 67 ms and 134 ms. For the [-ATR] close front vowels, the mean

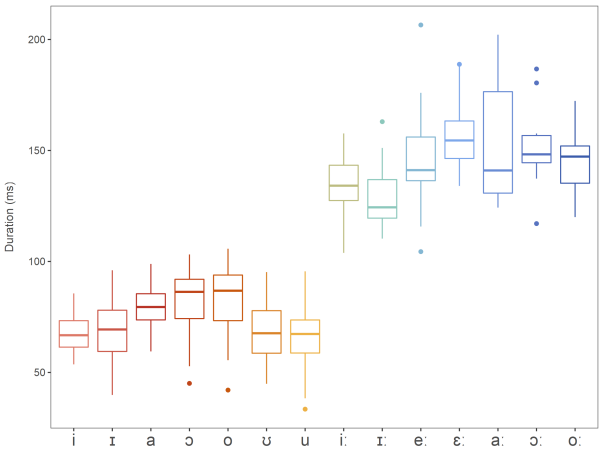


Figure 4: *Vowel duration of vowels in the first syllable.*

duration of the short vowel /ɪ/ and the long vowel /ɪː/ are 69 ms and 130 ms. For the open vowels, the duration of the short vowel /a/ and the long vowel /aː/ are 79 ms and 157 ms. For the [-ATR] mid-back vowels, the duration of the short vowel /ɔ/ and the long vowel /ɔː/ are 81 ms and 151 ms. For the [+ATR] mid-back vowels, the duration of the short vowel /o/ and the long vowel /oː/ are 82 ms and 145 ms. Across all vowel qualities, short vowels are on average 73 ms and long vowels are on average 146 ms, and therefore exactly twice as long.

The duration measures in the first syllable suggest that five vowel qualities exhibit a length contrast, including /i, iː/, /ɪ, ɪː/, /a, aː/, /o, oː/, and /ɔ, ɔː/, and phonological evidence across syllables indicates that /e, eː/ and /ɛ, ɛː/ also contrast in length. Furthermore, as is shown in Figure 4, the duration of the Kufo vowels displays a pattern where the closer vowels have a shorter duration compared to the more open ones, and this pattern is widely attested across languages.

It is likely that the close and near-close back vowels /u, ʊ/ also exhibit a length contrast, yet there is no direct supporting evidence in the current database. However, as can be observed from the results, the duration of /u/ and /ʊ/ are in a similar range to the durations for short vowels where long counterparts are established, so it is possible that these two vowel qualities both exhibit a length contrast if there is more lexical data available for examination, and some examples may occur in data other than disyllabic words.

5. Conclusion and discussion

The acoustic phonetic results of the present study provide supporting evidence for nine contrastive vowel qualitites (/i, ɪ, e, ɛ, a, ɔ, o, ʊ, u/) as well as a length contrast for five vowel qualities where this can directly be tested in the first syllable of disyllabic words (/i, iː/, /ɪ, ɪː/, /a, aː/, /ɔ, ɔː/, /o, oː/) and two vowel qualities with phonological evidence across syllables (/e, eː/, /ɛ, ɛː/). Whether the vowel qualities /u, ʊ/ also display a length contrast requires further phonetic evidence when more lexical data is available. This study has not examined the phonetic correlates of the seemingly non-contrastive mid central vowel [ə] due to its restricted contexts. Based on the results of the present study, a Kufo vowel inventory is presented in Table 4, with the mid central vowel marked in square brackets and proposed phonemes requiring further phonetic evidence accompanied by asterisks.

Table 4: *The Kufo vowel inventory.*

	front		central		back	
	short	long	short	long	short	long
close	/i/	/iː/			/u/	*/uː/
near-close	/ɪ/	/ɪː/			/ʊ/	*/ʊː/
close-mid	/e/	/eː/			/o/	/oː/
mid			[ə]			
open-mid	/ɛ/	/ɛː/			/ɔ/	/ɔː/
open	/a/	/aː/				

This study constitutes an acoustic phonetic investigation of the vowel inventory of the Kufo language for which the linguistic record was, until recently, extremely limited. The acoustic phonetic results and proposed vowel inventory lay the groundwork for future studies on the Kufo sound system. As more lexical data for the language becomes available, it will be possible to develop a more comprehensive database for quantitatively studying the Kufo vowel inventory, with sufficient data for statistical investigations to also be undertaken. Apart from having a perfect balance for all vowel phonemes of contrastive qualities and lengths, an ideal Kufo vowel database would consist of disyllabic CVCV words where the adjacent consonants are also controlled. One open question for Kufo is whether there is any presence of vowel harmony, given that there are four [+/-ATR] pairs that display an ATR-type contrast (/i, ɪ/, /e, ɛ/, /o, ɔ/, and /u, ʊ/) based on their F1 measurements, and vowel harmony is widely attested across languages with an ATR contrast [10]. While there is as yet no morphophonological evidence for this, the acoustic measurement results in this study offer a reference point for more closely examining the [+ATR] or [-ATR] status of vowels in other data, allowing for closer investigation of any possible vowel harmony. The status of the mid central vowel [ə] in Kufo also requires closer investigation, since it is describe as closely related to /a/ in previous studies, and it may be a reduced realisation of /a/ in specific morphological environments based on the data collected for the current study, especially considering how /a/ appears to have a wider range in F1 and F2 measurement compared to other vowel qualities. When more data become available in the future, /a/ will be closely examined regarding the relationship between its quality and duration.

The present study aids preparation for future studies on tonal patterns in the Kufo language. Though the Kadu languages are often hypothesised as tonal, their tone systems have never been systematically studied. The findings on contrastive vowel quality and length allow these characteristics to be controlled for in analyses focusing on tone, in order for the prosodic system of Kufo to be better understood.

The present findings also provide insights into the vowel systems of the Kadu language family more generally, whose phonemic inventories have not yet been comprehensively studied and phonetically examined. The languages of the Nuba Mountains in Sudan are extremely diverse but very understudied, and in some cases highly endangered, and there is much still to be learned about the linguistic structures and relationships in this region. More broadly, the present study contributes to the phonological studies of African languages based on phonetic data. While there are rich descriptions on interesting phonological phenomena for African languages, phonetic data is rarely brought to bear on discussions of African phonological systems.

6. Acknowledgements

A heartfelt *tido tikka* to Haroun Kafi for contributing his time and knowledge to this study, and to the Kufo community for their language. I wish to thank Rosey Billington for her encouragement, advice, and comments on this work. I am grateful to Nicholas Evans and Matthew Carroll for introducing me to Kufo, and Roger Blench and Abdalla Mongash for sharing their wordlist. My sincere thanks to the PARADISEC archive, especially Julia Miller, for safeguarding the data. Funding support from the ARC Centre of Excellence for the Dynamics of Language (Project ID: CE140100041) is gratefully acknowledged.

7. References

[1] Hammarström, H. & Forkel, R. & Haspelmath, M. & Bank, S., "Glottolog 4.6", Leipzig: Max Planck Institute for Evolutionary Anthropology, https://glottolog.org/resource/languoid/id/kang1288, 2022.

[2] Eberhard, D. M., Gary F. S., & Charles D. F., "Ethnologue: languages of the world", 25th edition, Dallas, Texas, SIL International, http://www.ethnologue.com, 2022.

[3] Greenberg, J., "Studies in African linguistic classification", New Haven, Compass Publishing Company, 1955.

[4] Schadeberg, T., "The classification of the Kadugli language group", in Nilo-Saharan: Proceedings of the First Nilo-Saharan Linguistics Colloquium, Leiden, September 8-10, pp.291-305, 1980.

[5] Hall, E. and Hall, M., "Kadugli-Krongo", Occasional Papers in the Study of Sudanese Languages, 9:57-67, 2004.

[6] Blench, R. M., "The Kadu languages and their affiliation: between Nilo-Saharan, Niger-Congo and Afro-Asiatic", in Insights into Nilo-Saharan Language, History and Culture, Proceedings of the 9th Nilo-Saharan Linguistics Colloquium, Institute of African and Asian Studies, University of Khartoum, 16-19 February 2004, 2006.

[7] Schadeberg T. C., "Comparative Kadu wordlists", in Afrikanistische Arbeitspapiere: Schriftenreihe des Kölner Instituts für Afrikanistik 40: 11-48, 1994.

[8] Blench, R. M. & Mongash, A., "The Kufo language of the Nuba Hills, Sudan" [unpublished manuscript], 2022.

[9] Evans, N. et al., "Developing an orthography on Kufa (Kadugli-Krongo): three axes of design", 15th Nilo-Saharan Linguistics Colloquium, 2021.

[10] Casali, R. F., "ATR harmony in African languages", Language and Linguistics Compass, 2(3):496-549, 2008.

[11] Guion, S. G., Post, M. W., & Payne, D. L., "Phonetic correlates of tongue root vowel contrasts in Maa", in Journal of Phonetics, 32:517-542, 2004.

[12] Billington, R., "Advanced Tongue Root in Lopit: Acoustic and ultrasound evidence", in Proceedings of the 15th Australasian International Speech Science & Technology Conference, pp.119-122, 2014.

[13] Hall, R. M. R. and Creider, C., "The fates of [+ATR] /a/ in Nilotic", in I. Maddieson and T.J. Hinnebusch [Eds], Language history and linguistic description in Africa, 45-54, African World Press, 1998.

[14] Remijsen, B., "Evidence for three-level vowel length in Ageer Dinka", in Above and Beyond the Segments, pp.246-260, 2014.

[15] Remijsen, B., Ayoker, O. G., & Jørgensen, S., "Tenary vowel length in Shilluk", in Phonology 36-1:91-125, Cambridge University Press, 2019.

[16] Carroll, M., J. (collector), "Field Methods ANU 2021 Kufa", Collection KCP1 at http://catalog.paradisec.org.au/collections/KCP1, 2021.

[17] Li , S. (collector), "Shubo Li Kufo Materials", Collection KCP2 at http://catalog.paradisec.org.au/collections/KCP2, 2022.

[18] Boersma, P. and Weenick, D., "Praat: Doing phonetics by computer" [computer program], www.praat.org, 2016.

[19] Wells, J. C., "Computer-coding the IPA: A proposed extension of SAMPA" [unpublished manuscript], 1995.

[20] Croot, K., & Taylor, B., "Criteria for acoustic-phonetic segmentation and word labelling in the Australian National Database of Spoken Language", Speech, Hearing and Language Research Centre, Macquarie University, 1995.

[21] Winkelmann, R., Harrington, J., & Jaensch, K., "EMU-SDMS: Advanced speech database management and analysis in R", in Computer Speech & Language, 45, 392-410, 2017.

[22] 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.

[23] R Core Team, "R: A language and environment for statistical computing" [computer program], https://www.R-project.org, Vienna, Austria, 2022.

[24] RStudio Team, "RStudio: Integrated development environment for R" [computer program], http://www.rstudio.com/, RStudio, PBC, Boston, MA, 2022.

Voice Quality of the Nasal Vowels in Chaoshan Chinese

Changhe Chen

Department of Linguistics, The University of Hong Kong
riverch8@connect.hku.hk

Abstract

Previous studies have shown that nasal/nasalized vowels can be produced with breathier voice, possibly due to speech enhancement or misperception [1-2, 20]. Thus, this study investigates whether the nasal vowels of Chaoshan Chinese are also breathier than their oral congeners, and if so, how creakiness of the tone /212/ interacts with the nasality and breathier voice.

Electroglottograph (EGG) data were collected from native speakers of Chaoshan Chinese. Contact Quotient (CQ) data of 10 speakers show that the nasal vowels of this language can also be breathier than the oral vowels, but such difference in voice quality is optional and speaker dependent. In the condition of tone /212/ with periodic voice, creaky voice of the tone may override the redundant breathiness of the nasal vowels. However, when tone /212/ is produced with aperiodic voice as observed in the oral vowels, data of HNR05 show that the nasal counterparts are more periodic; a hypothesis is given to explain it.

Index Terms: nasal vowel, breathy voice, creaky voice, speech enhancement, Chaoshan Chinese

1. Introduction

Previous studies have demonstrated that, apart from lowered velum, nasal/nasalized vowels can be produced with breathier voice [1, 2], distinct labial and/or lingual gestures [3-7], and different pharyngeal configurations [6, 7]. These constitute the "multidimensionality" of the nasal vowel articulation [1]. Since there is no comprehensive articulatory study on the nasal vowels in Sinitic languages, we are carrying out a project to investigate them. This study focuses on the voice quality of the nasal vowels in Chaoshan Chinese, a dialect of Southern Min spoken in Guangdong province of China.

The representative dialect of Chaoshan Chinese is spoken in Shantou [8]. Its sound inventory is presented in Table 1.

Table 1: *The sound inventory of Shantou dialect [8].*

Consonants	/p pʰ b t tʰ k kʰ g ʔ s z h ʦ ʦʰ m n ŋ l/
Vowels	/i ɛ a ɔ ɯ u ĩ ɛ̃ ã/
Diphthongs	/ia iɔ iu ai au ɔi ɔu ui uɛ ua ĩã ĩɔ ĩũ ãĩ ãũ ɔĩ ɔũ ũĩ ũɛ̃ ũã/
Triphthongs	/iau ĩãũ uai ũãĩ/
Tones	/33 55 51 35 212 22 2 5/

The IPA symbols and tone letters used here are based on the data collected for this study. The dipping tone /212/ can be produced with creaky voice when the F0 is low (Figure 1 shows two examples); the degree and type of creakiness are variable among speakers. Creakiness was also observed in [9]. Due to insufficient study on it, it is unknown whether the creakiness is a byproduct of low F0, like tone /213/ in Mandarin [10]. The phonemic nasal vowels in this language derived historically from oral vowels by deleting nasal codas or spontaneous nasalization of oral vowels [8]; and this is the basis of comparing the oral and nasal vowels. The nasal vowels do not have any excrescent nasal consonants [8].

Nasal vowels are produced with lowered velum, and such articulatory gesture introduces nasal formants and anti-formants, which result in broader F1 bandwidth, lower F1 amplitude, and higher spectral tilt [11-17]. These consequent acoustic features were also found in breathy voice resulting from less constricted glottis [2, 18-20]. Possibly due to misperception or enhancement, speakers use breathier voice in producing nasalized vowels of Yi languages [2] and nasal vowels of French [1]. The connection between breathiness (due to high airflow) and nasalization also manifests in the historical sound change in Chaoshan Chinese, where the oral vowels after high airflow consonants were nasalized, such as in words 怕 "scared" [pʰã^{212}], 虎 "tiger" [hɔ̃ũ^{51}], 鼻 "nose" [pʰĩ^{22}] and so on [21]. Thus, the first research question of this study is whether native speakers of Chaoshan Chinese use breathier voice in producing the nasal vowels. According to previous studies, Chaoshan Chinese does not have any phonemic breathy or creaky vowels, or breathy tones.

In articulation, breathy voice is generally produced with less glottal constriction than modal voice and with different degrees of noise dependent on the types of breathy voice [22]; thus, the contacting percentage of a glottal cycle is lower (lower contact quotient (CQ) value in the EGG measurement). In contrast, one common articulatory feature of creaky voice is greater glottal constriction (except "nonconstricted creak") [23]; thus, the CQ value is higher.

As mentioned above, tone /212/ is produced with creaky voice for some speakers and it can be aperiodic voice and/or have high damping. Lower spectral tilt [23-24] and possible narrower formant bandwidth [23, 25] of creaky voice are against the acoustic features of nasality and breathy voice. Thus, the creakiness may be a threat to the cues of nasality; it can also be a case of "overlap", though it is different from those in connected speech as discussed in [26]. As there is no study on nasal vowels with creaky/aperiodic voice, this study will investigate how speakers deal with it.

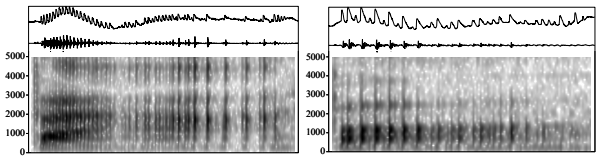


Figure 1: */pa^{212}/ produced by two speakers. The upper channel is the EGG signal.*

2. Method

2.1. Speakers

Data of 10 native speakers of Chaoshan Chinese were analyzed. These speakers were born in different cities of the Chaoshan region, namely Shantou, Chenghai, Chaozhou, Raoping, Jieyang and they are speaking the local dialects. All of them reported that they lived in Chaoshan till 18 years old, except one speaker from Raoping who moved to Guangzhou at 11 years old but still speaks Chaoshan Chinese at home. All of them can speak Mandarin and English as well, and some of them also speak Cantonese. None of them reported any speech and/or hearing disorders.

According to [8], there are only a few differences regarding the sound system among those dialects of Chaoshan Chinese. The consonants and tones are basically identical, while diphthongs /iɔ ĩɔ̃ iau ĩãũ/ in Shantou dialect are /iɛ ĩɛ̃ iɔu ĩɔ̃ũ/ in Chaozhou and Chenghai dialects.

2.2. Materials

The word list contained 45 (near) minimal pairs of oral and nasal vowels in (C)V syllables (C: consonant, V: monophthong/diphthong/triphthong) carrying high level tone /55/, mid-level tone /33/, low mid-level tone /22/, and low dipping tone /212/. However, the number of tokens with /22/ is small, so this tone is not included in the analysis. Table 2 presents some samples from the word list. The word list was randomized before presenting to the participants.

Table 2: *Sample syllables in the word list.*

Oral	Nasal
[i^{55}] 姨 “aunt”	[ĩ^{55}] 圓 “circle”
[pɛ^{55}] 爬 “to crawl”	[pɛ̃^{55}] 棚 “shed”
[pi^{33}] 碑 “stele”	[pĩ^{33}] 邊 “side”
[kɛ^{33}] 家 “home”	[kɛ̃^{33}] 羹 “thick soup”
[a^{33}] 亞 (亞洲) “Asia”	[ã^{33}] 搵 “to cover”
[i^{212}] 意 “intention”	[ĩ^{212}] 燕 “swallow”
[kɛ^{212}] 價 “price”	[kɛ̃^{212}] 頸 “neck”
[ka^{212}] 教 “to teach”	[kã^{212}] 酵 “to ferment”

2.3. Data collection

The EGG data were collected in a soundproof booth in 2021 by using the electroglottograph of Voce Vista (model: 7050A). Participants were instructed to produce three successive repetitions of each word in isolation with a pause between each token. EGG, ultrasound, lip movement and audio were recorded at the same time.

2.4. Data analysis

Since both nasality and breathy voice lead to an increase in spectral tilt, acoustic measures like H1*-H2*, H1*-A1* used in previous studies (e.g., [22]) to detect non-modal phonation in oral vowels become unreliable here. And as lower contact quotient (CQ) is related to an increase in spectral tilt [18], CQ should be measured to detect the breathy voice in nasal vowels; lower CQ indicates breathier voice. In addition, some speakers produced tone /212/ with aperiodic voice, where it is difficult to obtain reliable CQ values, so for those speakers HNR05 (harmonic-to-noise ratio under 500 Hz) was used to detect aperiodicity, where HNR values will be lower [23]. CQ and HNR05 were measured for the vocalic intervals by using EggWorks [27] and VoiceSauce [28] respectively. The standard percentage method is used in EggWorks, and the threshold percent is 25%. Data within each ninth of the token were averaged, and the resulting data (9 points for each token) were analyzed by using Smoothing Spline ANOVA (SSANOVA) to test the differences between the nasal and oral vowels. Shading around the curves indicates 95% confidence interval; overlap suggests no significant difference.

As EGG measures the vocal folds vibration at the glottis, it is assumed that different lingual gestures will have neglectable effect on the EGG data, so all vowels are pooled in the analysis. Some recordings were not used because of unfamiliar words, spontaneous nasalization of oral vowels, or clipping of the EGG signals. Near minimal pairs with onsets differing in the amount of airflow were also excluded (e.g., unaspirated stop/affricate vs. aspirated stop/affricate, stop vs. fricative). The numbers of (near) minimal pairs of oral and nasal vowels included in the analysis for each speaker are presented in Table 3.

Table 3: *The numbers of (near) minimal pairs of each participant included in the analysis.*

	Tone /55/	Tone /33/	Tone /212/
01M	24 pairs	42 pairs	23 pairs
02F	36 pairs	45 pairs	27 pairs
03M	24 pairs	30 pairs	18 pairs
04F	30 pairs	40 pairs	28 pairs
05F	24 pairs	39 pairs	27 pairs
06M	21 pairs	42 pairs	21 pairs
07M	21 pairs	25 pairs	12 pairs
08F	27 pairs	39 pairs	30 pairs
09M	21 pairs	33 pairs	19 pairs
10M	27 pairs	37 pairs	30 pairs

3. Results

Preliminary data analysis showed that voice quality in the nasal vowels is speaker and tone dependent. Thus, the results of each participant and tone are presented separately. For 03M, 05F, 06M and 08F, the third figures for tone /212/ present the SSANOVA results of HNR05 rather than CQ.

3.1. 01M (male speaker, aged 22 from Shantou)

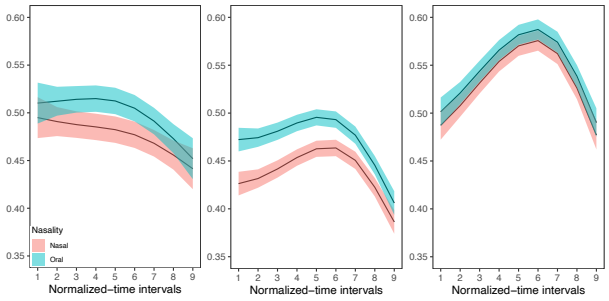


Figure 2: *SSANOVA results of CQ of /55 33 212/ (left to right).*

The nasal vowels have lower CQ when carrying tones /55/ (the middle part) and /33/, but there is no significant difference in the condition of tone /212/, which is produced with creaky voice.

3.2. 02F (female speaker, aged 29 from Chenghai)

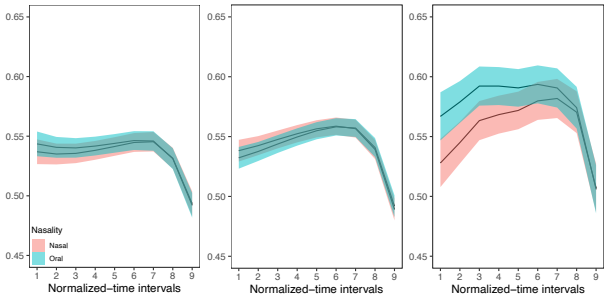


Figure 3: *SSANOVA results of CQ of /55 33 212/ (left to right).*

There is no significant difference between the nasal and oral vowels carrying tones /55/, /33/ and /212/ for this speaker. Tone /212/ is more variable.

3.3. 03M (male speaker, aged 28 from Shantou)

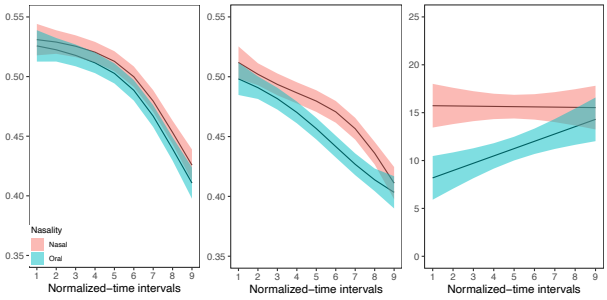


Figure 4: *SSANOVA results of CQ of /55 33/ (left and middle) and the result of HNR05 (dB) of /212/ (right).*

There is no significant difference between the oral and nasal vowels carrying tone /55/, and unexpectedly in the condition of tone /33/, the oral vowels have lower CQ than the nasal ones. HNR05 data show that from 1st to 6th interval the nasal vowels are more periodic than the oral vowels.

3.4. 04F (female speaker, aged 24 from Chenghai)

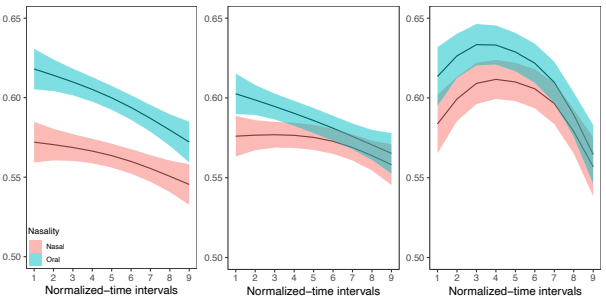


Figure 5: *SSANOVA results of CQ of /55 33 212/ (left to right).*

The nasal vowels have lower CQ than their oral congeners when carrying tones /55/ and /33/ (in the beginning), but there is no significant difference in the condition of tone /212/.

3.5. 05F (female speaker, aged 22 from Shantou)

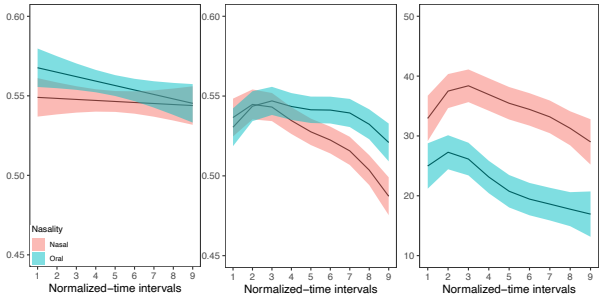


Figure 6: *SSANOVA results of CQ of /55 33/ (left and middle) and the result of HNR05 (dB) of /212/ (right).*

The nasal vowels carrying tone /33/ have lower CQ in the latter half. The nasal vowels with /212/ are more periodic than the oral vowels.

3.6. 06M (male speaker, aged 27 from Raoping)

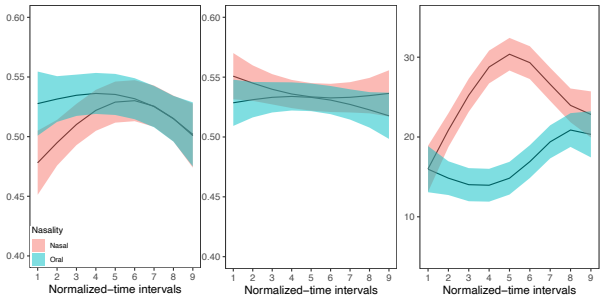


Figure 7: *SSANOVA results of CQ of /55 33/ (left and middle) and the result of HNR05 (dB) of /212/ (right).*

There is no significant difference between the nasal and oral vowels carrying tones /55/ and /33/. However, the nasal vowels carrying /212/ are more periodic than the oral vowels.

3.7. 07M (male speaker, aged 29 from Jieyang)

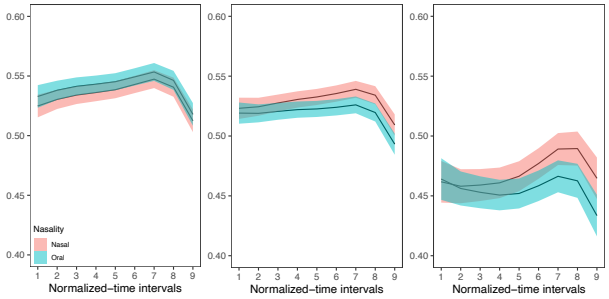


Figure 8: *SSANOVA results of CQ of /55 33 212/ (left to right).*

There is no significant difference between the nasal and oral vowels carrying tones /55/, /33/ and /212/. Tone /212/ is breathier than /55 33/.

3.8. 08F (female speaker, aged 22 from Shantou)

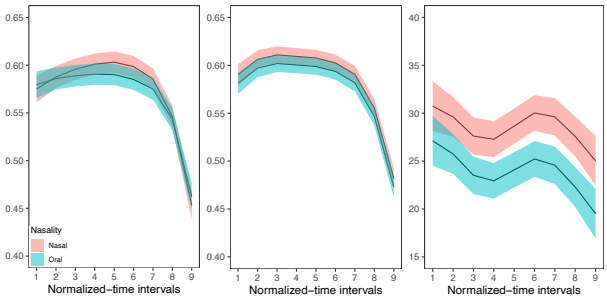


Figure 9: *SSANOVA results of CQ of /55 33/ (left and middle) and the result of HNR05 (dB) of /212/ (right).*

The nasal vowels carrying /212/ are more periodic than the oral vowels.

3.9. 09M (male speaker aged 29 from Chaozhou)

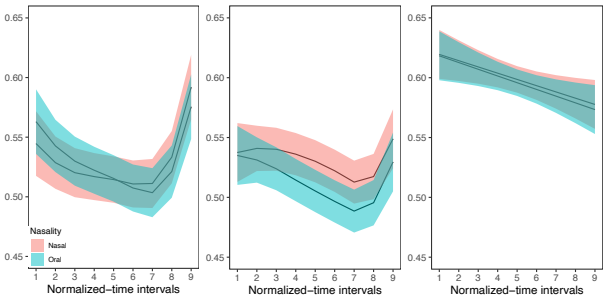


Figure 10: *SSANOVA results of CQ of /55 33 212/ (left to right).*

There is no significant difference between the nasal and oral vowels with tones /55/, /33/ and /212/ for this speaker.

3.10. 10M (male speaker aged 32 from Shantou)

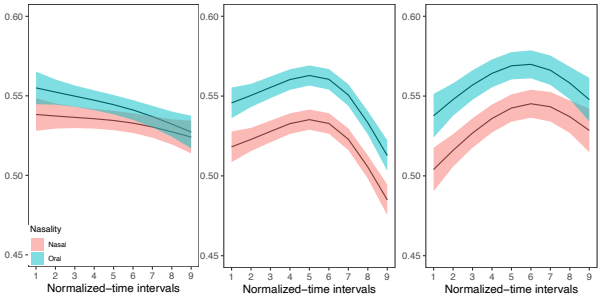


Figure 11: *SSANOVA results of CQ of /55 33 212/ (left to right).*

For this speaker, the nasal vowels have lower CQ than the oral vowels in the conditions of tones /33 212/.

4. Discussion and Conclusions

4.1. Voice quality of the nasal vowels

Seven out of ten speakers show lower CQ or higher HNR05 (for tone /212/) in the nasal vowels, suggesting breathier voice or more periodic voice. Due to the shared acoustic features of nasality and breathy voice, breathiness in nasal vowels is believed to enhance the percept of nasality [1-2, 20].

Regarding the breathier voice, this study demonstrates two properties of it. First, it is optional. Not all speakers use breathier voice in producing nasal vowels, and for those who use it, they do not always use it; for example, 05F uses it in the condition of tone /33/ but not or less in tone /55/. The nasal vowels of 03M are an exception. In the tone /33/ condition, the nasal vowels have higher CQ, contrary to the prediction and other speakers. The reason is unknown, but it also indicates that breathier voice is not necessary. Second, it is variable. There is no consistent pattern across speakers regarding when (the time) and where (the tone) it occurs. It can occur in the initial (04F), middle (01M) or latter (05F) part or the entire (04F and 10M) of the vowels. Also, speakers use breathier voice in different conditions of tones. For example, 04F uses it when the nasal vowels carry tones /55 33/ but not or less in tone /212/, while 10M uses it in tones /33 212/ but not or less in tone /55/. High variability of voice quality in nasal vowels resembles the use of rounding and sublingual cavity in producing English [ʃ], which lower the resonance frequency and enhance the differences between [s] and [ʃ] [14]. [29] showed variation in the amount of lip rounding and the size of sublingual cavity employed by speakers.

4.2. Nasal vowels with creaky voice

Tone /212/ with creaky voice is classified into two types. First, the voice is creaky but still periodic (such as 01M, 02F, 04F, 09M). In this case, the nasal vowels are not breathier than the oral congeners. For speakers who use breathier voice in tones /55 33/ (01M and 04F), creaky voice of /212/ may have overridden the enhancing gesture. It is possible that for these speakers the creaky voice is a defining feature of tone /212/, so it cannot be breathier. Second, tone /212/ is produced with aperiodic voice in the oral vowels (03M, 05F, 06M, 08F). Data of HNR05 show that the nasal counterparts generally are more periodic than the oral vowels. It means the nasality can override the aperiodicity of tone /212/.

One reviewer mentioned that the creaky voice may be produced with ventricular incursion and/or aryepiglottic constriction. It is conceivable since we can see from Figure 12 that in the framed part the vibrations of the vocal folds only appear in the EGG signal (the upper channel), and possibly the sound is muffled by the constriction above the glottis. I hypothesize that such articulatory gesture may hinder the airflow necessary for nasalization, and thus the nasal vowels tend to be produced with less constriction than the oral vowels. Further studies are needed to corroborate it.

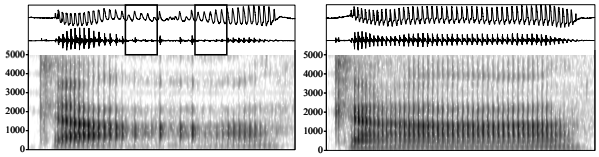


Figure 12: */ka^{212}/ (left) and /kã^{212}/ (right) of 06M.*

In sum, the nasal vowels of Chaoshan Chinese can be breathier than the oral counterparts, but such difference is optional and variable. Creaky voice may override breathiness, which is an enhancing and redundant feature, and the nasality can override the aperiodicity of tone /212/.

5. Acknowledgements

I'd like to thank the anonymous reviewers for their detailed and instructive comments.

6. References

[1] C. Carignan, "Covariation of nasalization, tongue height, and breathiness in the realization of F1 of Southern French nasal vowels", *Journal of Phonetics*, 63, pp. 87-105, 2017.

[2] M. Garellek, A. Ritchart, & J. Kuang, "Breathy voice during nasality: A cross-linguistic study", *Journal of Phonetics*, 59, pp. 110-121, 2016.

[3] R. Shosted, "Nasal vowels are not [+nasal] oral vowels", in *Romance Linguistics 2012: Selected papers from the 42nd Linguistic Symposium on Romance Languages*, J. Smith & T. Ihsane (Ed.), Amsterdam: Jon Benjamins, pp. 63-76, 2015.

[4] C. Carignan, "An acoustic and articulatory examination of the "oral" in "nasal": The oral articulations of French nasal vowels are not arbitrary", *Journal of phonetics*, 46, pp. 23- 33, 2014.

[5] R. Shosted, C. Carignan, & P. Rong, "Managing the distinctiveness of phonemic nasal vowels: Articulatory evidence from Hindi", *The Journal of the Acoustical Society of America*, vol. 131, no. 1, pp. 455-465, 2012.

[6] M. Barlaz, R. Shosted, M. Fu, & B. Sutton, "Oropharygneal articulation of phonemic and phonetic nasalization in Brazilian Portuguese", *Journal of Phonetics*, 71, pp. 81-97, 2018.

[7] C. Carignan, R. K. Shosted, M. Fu, Z. P. Liang, & 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", *Journal of phonetics*, 50, pp. 34-51, 2015.

[8] X. Li, *Chinese varieties in Guangdong province*. Guangzhou: Guangdong People's Publishing House, 1994. [in Chinese]

[9] Y. Hong, *A phonetic study of Chaozhou Chinese*, Doctoral dissertation, HKUST, 2013.

[10] J. Kuang, "Covariation between voice quality and pitch: Revisiting the case of Mandarin creaky voice", *The Journal of the Acoustical Society of America*, vol. 142, no. 3, pp. 1693-1706, 2017.

[11] M. Y. Chen, "Acoustic correlates of English and French nasalized vowels", *The Journal of the Acoustical Society of America*, vol. 102, no. 4, pp. 2360-2370, 1997.

[12] O. Fujimura, & J. Lindqvist, "Sweep-tone measurements of vocal ‐ tract characteristics", *The Journal of the Acoustical Society of America*, vol. 49, no. 2B, pp. 541-558, 1971.

[13] S. Hawkins, & K. N. Stevens, "Acoustic and perceptual correlates of the non-nasal - nasal distinction for vowels", *The Journal of the Acoustical Society of America*, vol. 77, no. 4, pp. 1560-1575, 1985.

[14] K. Johnson, *Acoustic and auditory phonetics*. John Wiley & Sons, 2012.

[15] K. N. Stevens, *Acoustic phonetics*. MIT press, 2000.

[16] K. N. Stevens, G. Fant, & S. Hawkins, "Some Acoustical and Perceptual Correlates of Nasal Vowels", in *In Honor of Ilse Lehiste*, R. Channon & L. Shockey (Eds.), De Gruyter Mouton, pp. 241- 254, 1987.

[17] W. Styler, "On the acoustical features of vowel nasality in English and French", *The Journal of the Acoustical Society of America*, vol. 142, no. 4, pp. 2469–2482, 2017.

[18] E. Holmberg, R. Hillman, J. Perkell, P. Guiod, & S. Goldman, "Comparisons among aerodynamic, electroglottographic, and acoustic spectral measures of female voice", *Journal of Speech and Hearing Research*, vol. 38, pp. 1212-1223, 1995.

[19] H. M. Hanson, "Glottal characteristics of female speakers: Acoustic correlates", *Journal of the Acoustical Society of America*, vol. 101, pp. 455-481, 1997.

[20] S. J. Keyser, & K. N. Stevens, "Enhancement and Overlap in the Speech Chain", *Language*, vol. 82, no. 1, pp. 33-63, 2006.

[21] K.-Y. Chang, "Nasalization of Nasal Finals in Chinese", *Studies in Language and Linguistics*, vol. 32, no. 2, pp. 17-28, 2012. [in Chinese]

[22] J. Tian, & J. Kuang, "The phonetic properties of the non-modal phonation in Shanghainese", *Journal of the International Phonetic Association*, pp. 1–27, 2019.

[23] P. Keating, M. Garellek, & J. Kreiman, "Acoustic properties of different kinds of creaky voice", in *Proceedings of the 18th International Congress of Phonetic Science*, Glasgow, UK, 2015, pp. 0821.1-0821.5.

[24] M. Garellek, "The phonetics of voice", *The Routledge handbook of phonetics*, W. F. Katz and P. F. Assmann (Ed.), London and New York: Routledge, pp. 75-106, 2019.

[25] C. Gobl, "A preliminary study of acoustic voice quality correlates", *STL-QPSR*, vol. 4, pp. 9-21, 1989.

[26] K. N. Stevens, & S. J. Keyser, "Quantal theory, enhancement and overlap", *Journal of Phonetics*, vol. 38, pp. 10-19, 2010.

[27] H. Tehrani, "EGGWorks: A Program for Automated Analysis of EGG Signals", available online: https://appsobabble.com/functions/EGGWorks.aspx (accessed in March 2022).

[28] Y. Shue, P. Keating, C. Vicenik and K. Yu, "VoiceSauce: a program for voice analysis," in *Proceedings of the 17th International Congress of Phonetic Sciences*, Hong Kong, 2011, pp. 1846–1849.

[29] M. Proctor, C. Shadle, & K. Iskarous, "An MRI study of vocalic context effects and lip rounding in the production of English sibilants", in *Proceedings of the 11th Australasian International Conference on Speech Science and Technology*, Auckland, New Zealand, 2006, pp. 307-312.

The influence of pitch and speaker sex on the identification of creaky voice by female listeners

Hannah White, Andy Gibson, Joshua Penney, Anita Szakay, Felicity Cox

Department of Linguistics, Macquarie University

hannah.white2@hdr.mq.edu.au; andy.gibson, joshua.penney, anita.szakay, felicity.cox@mq.edu.au

Abstract

Past work has raised questions about how creaky voice quality is identified in different voices, particularly whether greater pitch differences between modal and creaky voice facilitates creak identification or whether social expectations bias identification in certain voices. While the role of pitch and speaker sex in creak identification has been investigated, results have been equivocal. In this study, we used highly controlled stimuli to examine the extent to which pitch and/or speaker sex affect the identification of creak. A study of 130 Australian English-speaking female listeners found that pitch and speaker sex interacted in listeners' perception of creaky voice.

Index Terms: creaky voice, perception, speaker sex, pitch, Australian English

1. Introduction

Creaky voice is a non-modal voice quality, typically perceived to be rough, pulse-like and low in pitch [1, 2, 3, 4]. Acoustic analyses have shown there are many different phonetic realisations of creaky voice [2]; however perception research suggests that low pitch is a salient cue to the perception of all realisations despite their phonetic differences [5], and that the presence of creak can lead listeners to assign lower pitch ratings to these utterances compared to fully modal utterances [6].

In media reports and popular culture, creak is generally associated with young women's speech and is attached predominantly to negative connotations [7, 8, 9]. These negative associations are likely linked to the prominent use of creak by celebrities such as Kim Kardashian, Britney Spears and Paris Hilton [10]. Some linguistic perception research has corroborated the negative sentiments seen in media headlines. Such studies have found that voices with creak are rated as less competent, less educated, less trustworthy, less hirable and lower in solidarity than modal voices [11, 12]. These findings were especially strong in descriptions of female speakers compared to males [11, 12]. However, other studies have found more positive meanings associated with creak. Yuasa [13] found that male and female listeners rated American female speakers with creak as sounding casual, educated and genuine, and Gobl and Ní Chasaide [14] found that Irish listeners rated speakers with creak as sounding relaxed, intimate and unafraid.

In production research, many studies have suggested that creak is more prevalent in female speech, especially in American English [13, 15, 16, 17, 18]. However, Dallaston and Docherty [19] found that the majority of research in this area has been conducted in the United States, and that studies disproportionately focus on young women; therefore, more work is needed in order to empirically confirm whether creak is more prevalent in female speech than male speech. Studies from the United Kingdom have found creak to be more prevalent in the speech of men [20, 21, 22], as did a recent study of Australian English (AusE) [23], and a 2016 study of American English suggests no difference in creak prevalence between females and males [24]. Although the present study focuses solely on female listeners, future analyses will also include male listeners.

We know that creak is a feature used by both males and females in speech [16, 20, 24]; however, the general association of creaky voice with women in the wider population stands [7, 8, 9, 10]. This raises the question of whether listeners are biased to hear creak in female speech compared to male speech. Davidson [25] addressed this question, by investigating whether the identification of creak (i.e. whether a listener decides creak is present or not) is influenced more by listener expectations or by acoustic properties of the voice. Two male and two female speakers were used in her study. Within each sex one speaker had relatively high and one had relatively low average f0 [25]. The high male voice and low female voice were matched for average pitch as much as possible using natural speech. It was hypothesised that high identification of creak in female speech compared to male speech, regardless of pitch, could be explained by a bias for identifying creak in females due to listener expectations. However, it was also hypothesised that the difference in f0 between modal voice (typically higher for females) and creak (typically low for all speakers) would make creak more salient in female voices. Under this hypothesis, it was proposed that creak would be most identifiable in the high female voice, equally next identifiable in the low female and high male voices (as they were matched for f0) and least identifiable in the low male voice. Two experiments were run in which listeners were asked whether there was creak in the stimuli they were played. The second experiment used low female and high male voices that were more closely matched than in the first version. The results showed that listeners consistently false alarmed in the low male modal condition, i.e. they identified creak when it was not present. However, conflicting results across experiment versions meant it was not possible to attribute findings to any proposed hypothesis. In addition, due to the use of natural speech as stimuli, the high male and low female conditions could not be exactly matched for f0, making it impossible to unpack the influence of pitch and speaker sex on creak identification.

1.1. Research question and expectations

Our motivation is to disentangle whether pitch or speaker sex has greater influence on creak identification by using stimuli manipulated to ensure maximum control over speaker f0 and creaky voice. An experiment was designed to provide listeners with a highly controlled set of two-word noun phrases that varied according to speaker sex (female vs male), voice quality

(fully modal vs modal plus creaky component), and f0. Listeners were asked if they identified creak when presented with two types of stimuli: creaky, which contained creak, and modal, which did not contain creak (see Section 2.3). Source recordings from a male and female speaker were manipulated for f0 and presence of creak at the end of the phrase (i.e. creaky component). This resulted in four pitch conditions: low male, mid male, mid female, and high female. Importantly, mid male and mid female conditions had identical f0 contours, and differed only in source speaker sex. The process for creating the stimuli is described in Section 2.1.

We proposed three possible influences on identification of creaky voice (the first two from [25]):

1. Pitch: the difference in pitch between a speaker's modal voice and creaky voice has the greatest influence on accurate creak identification.
2. Speaker sex: the difference in the source speaker sex has the greatest influence on accurate creak identification.
3. Pitch and sex: pitch differences between modal voice and creaky voice and source speaker sex interact in creak identification.

In Table 1, we present our predictions for each scenario.

Influence	Creak condition	Modal condition
Pitch	hi-f > mid-f = mid-m > lo-m	hi-f > mid-f = mid-m > lo-m
Sex	hi-f = mid-f > mid-m = lo-m	lo-m = mid-m > mid-f = hi-f
Pitch & sex	hi-f > mid-f > mid-m > lo-m	hi-f > mid-m > mid-f > lo-m

Table 1: *Table of expectations of listener accuracy in creak identification for each proposed influence. hi-f = high female condition, mid-f = mid female condition, mid-m = mid male condition and lo-m = low male condition.*

If the pitch of the stimuli has the greatest influence on creak identification, we would expect creak to be more noticeable in the creak condition when the modal f0 is highest due to a clear difference in f0 between the modal and creak components in the stimuli. This would lead us to expect the highest creak identification accuracy in the high female pitch condition, followed by the mid female and mid male conditions (which have identical f0s), and lowest accuracy in the low male condition. We would expect the same patterning of results in the modal condition because the absence of creak would be most noticeable when modal f0 is highest. In the modal condition there is no substantial drop in f0 to trigger a creak response.

If speaker sex has the greatest influence on creak identification, we would expect creak identification accuracy patterns to be consistent within source speaker sexes in both creaky and modal conditions but different between the source sex conditions. In the creaky condition, we would expect listeners to be most accurate for the female voices and least accurate for the male voices regardless of f0 because of a bias towards hearing creak in female voices. In the modal condition, we propose that listeners would have lowest accuracy in not identifying creak in the female voice conditions due to this bias leading them to false alarm (i.e. identify creak for the female source stimuli when it is not present).

If pitch and speaker sex are interacting in creak identification, we propose that pitch would be a stronger influence on creak decisions in the most extreme pitch conditions (low male and high female). This means we would expect accuracy to be highest in identifying creak when it's present and not identifying creak when it's not present in the high female condition and lowest in the low male condition. In the mid conditions when f0 is the same regardless of speaker sex, we propose that speaker sex will mediate creak decisions. We would expect listeners to identify creak more in the mid female voice than the mid male voice in both the creaky and modal conditions due to the bias to identify creak in female voices. This would result in higher accuracy to creaky tokens and lower accuracy to modal tokens in the mid female condition compared to the mid male condition.

2. Methods

2.1. Stimuli

The stimuli were produced by one female (22 y.o.) and one male (28 y.o.), siblings, who are both native speakers of AusE. Each speaker recorded 30 two-word phrases. The phrases were chosen with reference to a list of adjective-noun pair bigrams extracted from ONZE [26] along with counts of how frequently each occurred in the corpus. Twenty frequent bigrams, including *huge pain*, *free time*, *large farm* and *main store*, were selected. A further 10 were created by pairing high frequency adjectives and high frequency nouns from the ONZE list, e.g. *warm tea* and *brown shoe*. All adjectives and nouns were monosyllabic and all contained long vowels and voiced codas in order to optimise pitch information available to the listeners throughout the duration of the stimuli.

Stimuli materials were recorded using a Sennheiser 416 microphone and Universal Audio Apollo Quad interface with a preamp at 44.1 kHz sampling rate into Logic Pro on a MacBook Pro computer. Speakers were asked to produce each phrase in modal voice with neutral intonation. In order to ensure the degree of pitch manipulation was similar across conditions, speakers were instructed to aim for target f0s. The female speaker was asked to produce the phrases at a target f0 of 175 Hz (mean across stimuli = 169 Hz, range = 163–177 Hz) and the male speaker was asked to aim for an f0 of 125 Hz (mean across stimuli = 122, range = 115–125 Hz). The male speaker is musically trained and coached the female speaker in producing the stimuli. Target f0s were approximated with reference to the the musical notes F3 for the female and B2 for the male. These are close approximates to 175 Hz and 125 Hz respectively. These targets were determined so the starting points of the f0 manipulations would be proportionate to the source f0 for each condition (discussed below).

F0 was manipulated in Praat [27] to create both modal and creaky pairs of each adjective-noun phrase (with creak in the final 40% of the noun in the creak condition). While we investigated various methods for creating creaky voice in our stimuli, we determined that manipulating f0 produced the most natural sounding creak. Firstly, two different pitch conditions were created per source voice (low male, mid male; mid female and high female). All the different pitch condition f0 manipulations were calculated proportionately. In lower pitch conditions (mid female and low male), f0 at the start of the adjective (first word in each phrase) was manipulated to be 85.6% of the source target. In the higher pitch conditions (high female and mid male), f0 at the start of the adjective was manipulated to be 120% of the source target. As a result, start f0 for the high female condition

was 210 Hz (120% of 175 Hz) and the start f0 for the low male condition was 107 Hz (85.6% of 125 Hz). These f0 values were chosen to reflect the mean f0s of 18–29 year old AusE speaking females and males found by [28]. F0 start for the mid male and mid female was 150 Hz.

All phrases were manipulated to have a gradual declination. In all conditions, the f0 at the end of the adjective/start of the noun rhyme was 93.3% of the f0 at the start of the adjective. In the modal condition, nouns (the second word in each phrase) were manipulated to have a gradual declination following from the adjective. The f0 at the end of the noun was 80% of the f0 at the end of the adjective/start of the noun rhyme. Mean f0s in the modal condition were 97 Hz in the low male condition, 135 Hz in the mid male and female conditions and 190 Hz in the high female condition. In the creaky condition, nouns were manipulated to have identical creaky voice starting 40% through the rhyme as follows: 70 Hz at 40%, 60 Hz at 50%, 50 Hz at 65%, 70 Hz at 75%, 60 Hz at 90% and 50 Hz at 100% of the rhyme duration. Manipulating f0 in this way created the percept of prototypical creaky voice with low and irregular f0 [2].

These manipulations resulted in 240 stimuli (30 phrases x 8 conditions: 2 x voice quality conditions - modal vs creak x 4 pitch conditions - low male, mid male, mid female and high female). Table 2 shows the f0 manipulations for each of the conditions. It shows that in the mid conditions, f0 values were identical for male and female source recordings.

Condition	Adjective		Noun	
	Start	Rhyme end	Rhyme start	Rhyme end
Modal lo-m	107 Hz	100 Hz	100 Hz	80 Hz
Modal mid-m	150 Hz	140 Hz	140 Hz	112 Hz
Modal mid-f	150 Hz	140 Hz	140 Hz	112 Hz
Modal hi-f	210 Hz	196 Hz	196 Hz	157 Hz
Creaky lo-m	107 Hz	100 Hz	100 Hz	creak
Creaky mid-m	150 Hz	140 Hz	140 Hz	creak
Creaky mid-f	150 Hz	140 Hz	140 Hz	creak
Creaky hi-f	210 Hz	196 Hz	196 Hz	creak

Table 2: *F0 manipulations of adjective and noun pairs by condition.*

2.2. Participants

Listeners were 130 AusE-speaking females who reported completing all of their schooling in Australia and no history of hearing loss. The mean listener age was 21 (range: 17–55) and they were all undergraduate students of linguistics or psychology. Listeners were compensated with either course credit or a $20 supermarket voucher for their time.

2.3. Procedure

The experiment was built in PsychoPy [29] and run online via Pavlovia.org. Participants were instructed to be seated in a quiet environment wearing headphones. Prior to starting the task, participants were provided with examples of creaky voice in male and female AusE voices, different to those used in the experiment. Examples were natural creak but followed the same structure as the stimuli (adjective-noun bigrams with creak on the noun). Additionally, participants heard eight practice items (one in each condition), in different male and female AusE voices. No feedback was provided during the practice. Participants then proceeded to the main experiment. For each trial participants were presented with a fixation cross for 500 ms, followed by the orthography of the phrase. After another 500 ms, the sound file played. Participants were asked to respond with the M and Z keys on their keyboard for whether they heard creak in the phrase or not, as quickly and as accurately as possible. Each participant was presented with all 240 stimuli across eight pseudo-randomised blocks of 30. Ten second breaks were provided between blocks.

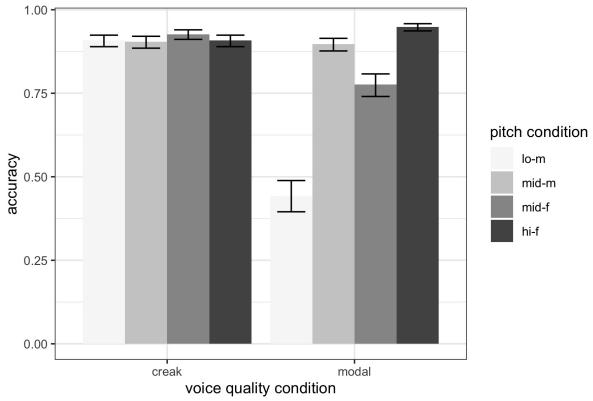


Figure 1: *Predicted accuracy results of generalised linear mixed effects model of each pitch condition (creaky voice condition and modal voice quality condition). Higher accuracy in the creak condition means more "creak" responses while higher accuracy in the modal condition means more "no creak" responses.*

3. Results

We limited our analysis to responses given after the offset of the adjective as it was the noun that contained the variable of interest (i.e. presence or absence of creak). We excluded responses that were quicker than 210 ms from the offset of the adjective or if they exceeded two standard deviations either side of the individual participant's mean response time. This resulted in the exclusion of 6.4% of the data leaving us with 29,213 data points for analysis.

A generalised linear mixed effects model was run using the lme4 package [30] in R [31] to investigate how listeners' accuracy at identifying creak presence or absence was affected by pitch and speaker sex in the different voice quality conditions (i.e. creaky and modal). Accuracy was included as the dependent variable and the independent variables were an interaction between four-level factor pitch condition (low male, mid male, mid female or high female) and two-level factor voice quality condition (creaky or modal), and scaled trial order with listener as a random intercept. We found a significant effect of trial order with listeners becoming more accurate throughout the duration of the experiment. The interaction between pitch condition and voice quality condition was also significant; model predictions are presented in Figure 1.

We conducted *post hoc* pairwise comparisons using the emmeans package [32] in R to further explore the interaction between pitch and voice quality. Results showed that in the creak condition, listeners were significantly more accurate at identifying creak presence in the mid female condition than in any other pitch conditions (low male: $p<0.05$; mid male: $p<0.01$; high

female: $p<0.05$). There were no other significant differences in the creak condition. In the modal condition, all pitch conditions were significantly different from each other (all $p<0.001$). Listeners were most accurate at identifying creak absence in the high female condition, followed by the mid male condition, the mid female condition and finally the low male condition. That is, they identified creak when it was not present most often in the low male stimuli and least often in the high female stimuli.

4. Discussion

As in previous literature [25, 33], our results show that listeners frequently false alarmed in the low male condition. In other words, they were more likely to identify creak in the low male condition when it was not present than for other conditions. However, we also see that the mid female condition also triggered false alarms when compared to the mid male and high female condition. Interestingly, listeners were significantly more accurate in identifying lack of creaky voice in the mid male condition despite the f0 manipulations being identical between the mid male and mid female conditions. These results point to listeners using both pitch and speaker sex as cues to identifying creaky voice: low pitch is the dominant cue triggering creak responses in the extreme pitch conditions regardless of speaker sex (i.e. in the modal condition the low male triggered inaccurate responses, but the high female triggered accurate responses). However, when the f0 is the same but speaker sex is different, the female source voice triggers creak responses.

Responses were at near ceiling accuracy in the creak condition. This tells us that listeners in this study were very good at identifying creak when it was present. Accuracy was significantly higher in the mid female creak condition compared with the other pitch conditions. We did not predict in any scenario that mid female stimuli would be rated more accurately than high female stimuli. We suggest that this may be due to a boost in the likelihood of identifying creak when pitch is low given a listener's expectations for sex.

Listeners may be making their judgments at least partially on the basis of the speaker's modal voice pitch (in this case, the adjective of the pair). This is highlighted in a comment left by one of the participants when asked if they noticed anything about the voices in the study:

> "The tone of the first word kind of gave away whether the second word would have creaky voice or not."

This is despite the adjective f0s being identical between the creaky and modal conditions. We suggest that, overall, when listeners hear a voice that is 'low for a female', or 'low for a male', they are more inclined to respond that they hear creak. In the creak condition, this would result in higher accuracy for the voices that sound low in the context of the speaker's sex (mid female and low male). In the modal condition, this boost would lead to false alarms for these speakers.

Even if listeners' judgments are affected by the pitch of the first word, it is clear that they are also sensitive to the creak manipulation. When listeners hear the creaky versions of the relatively high voices (high female and mid male), the difference in pitch between the initial higher modal voice and low creak leads to high rates of correct positive identifications of creak. In the modal condition, the absence of creak for these voices that sound relatively high given their sex is more obvious due to the lack of any dramatic drop in pitch. In future work, we will look at the reaction times to assess these possibilities. We may see faster responses where listeners' judgment relies on the pitch of the adjective.

If there is an overall boost in creak responses to the relatively low voices, we would expect to see greater accuracy to the low male voice in the creak condition. This may be absent in the present accuracy results due to the near ceiling effect. An analysis of reaction times may provide further clarity.

An alternative explanation for this over-identification of creak in the low male and mid female conditions could be based in experiences of hearing creak more often in low-pitched male and low-pitched female voices. A preliminary analysis of speaker creak prevalence by mean modal voice f0 suggests that creak may be more prevalent in male and female voices with lower mean modal f0s. The design of the present perception study does not enable us to determine whether the results in the modal condition are related to the association of creak with low pitch or the association of low-pitched voices using more creaky voice. Future research is needed to explore this from both a production and perception perspective.

It is possible that listener age impacts creak identification. Although the age range of listeners in the present study was quite large, the vast majority of listeners were less than 25 years old. The effects of listener age on creak identification would be an interesting area for future work.

5. Conclusions

This study set out to test how pitch and speaker sex influence female AusE-speaking females' identification of creaky voice. Results suggest that both pitch and speaker sex have an interrelated influence on creak identification by these listeners: for both female and male speakers, creak was over-identified in the lower-pitched condition. While previous work has not found solid evidence for sex and pitch influences [25], through the use of highly controlled manipulated stimuli, we have been able to tease apart these cues.

Findings from this study have implications for studies on the prevalence of creaky voice in speech. Many of these studies rely on manual annotation for identifying creak [19], which, as well as being a time-consuming and labour-intensive process, may lead to the over-identification of creak in lower-pitched male and female voices. There is a lot of recent research being conducted in the area of automatic creak detection methods, which may mitigate the biases of manual creak identification [19, 34, 35, 36, 37, 38].

In our future work we plan to investigate how male listener creaky voice identification compares to that of females. We also plan to explore whether reaction time data can shed further light on the role that speaker sex and pitch play in listeners' identification of creak.

6. Acknowledgements

This research was supported by a Macquarie University Research Excellence Scholarship to the first author and by Australian Research Council Grant DP190102164 and Australian Research Council Future Fellowship Grant FT180100462 to the fifth author. Thanks to members of the MQ phonetics lab, NZILBB lab and VUW Department of Linguistics for helpful discussions on this work.

7. References

[1] M. Garellek, "The phonetics of voice," in *The Routledge Handbook of Phonetics*, W. Katz and P. Assmann, Eds. Routledge,

2019.

[2] P. Keating, M. Garellek, and J. Kreiman, "Acoustic properties of different kinds of creaky voice," in *Proceedings of the 18th International Congress of Phonetic Sciences*. the University of Glasgow, 2015, Conference Proceedings, pp. 821.1–5.

[3] J. Peña, L. Davidson, and S. Orosco, "The independence of phrasal creak and segmental glottalization in American English," *JASA Express Letters*, vol. 1, p. 075205, 2021.

[4] L. Redi and S. Shattuck-Hufnagel, "Variation in the realization of glottalization in normal speakers," *Journal of Phonetics*, vol. 29, no. 4, pp. 407–429, 2001.

[5] L. Davidson, "Perceptual coherence of creaky voice qualities," in *Proceedings of the 19th International Congress of Phonetic Sciences*, S. Calhoun, P. Escudero, M. Tabain, and P. Warren, Eds. Australasian Speech Science and Technology Association Inc, 2019, pp. 196.1–5.

[6] ——, "Contributions of modal and creaky voice to the perception of habitual pitch," *Language*, vol. 96, no. 1, pp. e22–e37, 2020.

[7] B. Ralston, "Vocal fry is not okay, eh? - the listener," *The New Zealand Listener*, 2012. [Online]. Available: https://www.noted.co.nz/life/life-in-nz/vocal-fry-is-not-okay-eh/

[8] M. M. Weber. (2017) Top five most annoying vocal habits. [Online]. Available: https://www.voiceempowerment.com/voice-empowerment-blog/2017/5/1/ten-most-annoying-vocal-habits-or-5

[9] N. Wolf, "Young women, give up the vocal fry and reclaim your strong female voice," *The Guardian*, 2015. [Online]. Available: https://www.theguardian.com/commentisfree/2015/jul/24/vocal-fry-strong-female-voice

[10] A. Croffey, "Vocal fry: Women changing voices to sound "creaky" like Kim Kardashian," *The Sydney Morning Herald*, 2016. [Online]. Available: https://www.smh.com.au/lifestyle/vocal-fry-women-changing-voices-to-sound-creaky-like-kim-kardashian-20160428-gogwvo.html

[11] R. C. Anderson, C. A. Klofstad, W. J. Mayew, and M. Venkatachalam, "Vocal fry may undermine the success of young women in the labor market," *PLoS ONE*, vol. 9, no. 5, p. e97506, 2014.

[12] J. Pittam, "Listeners' evaluations of voice quality in Australian English speakers," *Language and Speech*, vol. 30, no. 2, pp. 99–113, 1987.

[13] I. P. Yuasa, "Creaky Voice: A New Feminine Voice Quality for Young Urban-Oriented Upwardly Mobile American Women?" *American Speech*, vol. 85, no. 3, pp. 315–337, 2010.

[14] C. Gobl and A. Ní Chasaide, "The role of voice quality in communicating emotion, mood and attitude," *Speech Communication*, vol. 40, pp. 189–212, 2003.

[15] N. B. Abdelli-Beruh, L. Wolk, and D. Slavin, "Prevalence of vocal fry in young adult male American English speakers," *Journal of Voice*, vol. 28, no. 2, pp. 185–190, 2014.

[16] S. Melvin and C. G. Clopper, "Gender variation in creaky voice and fundamental frequency," in *Proceedings of the 18th International Congress of Phonetic Sciences*, T. S. C. for ICPhS 2015, Ed. the University of Glasgow, 2015, pp. 1–5 (Paper number 320). [Online]. Available: https://www.internationalphoneticassociation.org/icphs-proceedings/ICPhS2015/Papers/ICPHS0320.pdf

[17] R. J. Podesva, "Gender and the social meaning of non-modal phonation types," *Annual Meeting of the Berkeley Linguistics Society*, vol. 37, no. 1, pp. 427–448, 2013.

[18] L. Wolk, N. B. Abdelli-Beruh, and D. Slavin, "Habitual use of vocal fry in young adult female speakers," *Journal of Voice*, vol. 26, no. 3, pp. e111–e116, 2012.

[19] K. Dallaston and G. Docherty, "The quantitative prevalence of creaky voice (vocal fry) in varieties of English: A systematic review of the literature," *PloS ONE*, vol. 15, no. 3, p. e0229960, 2020.

[20] B. Gittelson, A. Leemann, and F. Tomaschek, "Using Crowd-Sourced Speech Data to Study Socially Constrained Variation in Nonmodal Phonation," *Frontiers in Artificial Intelligence*, vol. 3, p. 565682, 2021.

[21] C. Henton and A. Bladon, "Creak as a sociophonetic marker," in *Language, speech, and mind: Studies in honour of Victoria A. Fromkin*, L. M. Hyman and C. N. Li, Eds. Routledge, 1988, pp. 3–29.

[22] J. Stuart-Smith, "Glasgow: Accent and voice quality," in *Urban voices: Accent studies in the British Isles*, P. Foulkes and G. Docherty, Eds. Arnold, 1999, pp. 203–222.

[23] D. Loakes and A. Gregory, "Voice quality in Australian English," *JASA Express Letters*, vol. 2, no. 8, p. 085201, 2022.

[24] N. Abdelli-Beruh, T. Drugman, and R. H. Red Owl, "Occurence frequencies of acoustic patterns of vocal fry in American English speakers," *Journal of Voice*, vol. 30, no. 6, pp. 759.e711–759.e720, 2016.

[25] L. Davidson, "The effects of pitch, gender, and prosodic context on the identification of creaky voice," *Phonetica*, vol. 76, no. 4, pp. 235–262, 2019.

[26] E. Gordon, M. Maclagan, and J. Hay, "The ONZE Corpus," in *Creating and digitizing language corpora*, J. C. Beal, K. P. Corrigan, and H. L. Moisl, Eds. Palgrave Macmillan, 2007, vol. 2, ch. 4, pp. 82–104.

[27] P. Boersma and D. Weenink, "Praat: Doing phonetics by computer," 2018, [Computer program]. [Online]. Available: http://www.praat.org/

[28] 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.e1–434.e15, 2022.

[29] J. Peirce, J. R. Gray, S. Simpson, M. MacAskill, R. Höchenberger, H. Sogo, E. Kastman, and J. K. Lindeløv, "Psychopy2: Experiments in behavior made easy," *Behavior Research Methods*, vol. 51, pp. 195–203, 2019.

[30] D. Bates, M. Maechler, B. Bolker, and S. Walker, "Fitting linear mixed-effects models using lme4," *Journal of Statistical Software*, vol. 67, no. 1, pp. 1–48, 2015.

[31] R Core Team, "R: A language and environment for statistical computing," 2021. [Online]. Available: https://www.R-project.org/

[32] R. Lenth, "emmeans: Estimated marginal means, aka least-squares means," 2018. [Online]. Available: https://CRAN.R-project.org/package=emmeans

[33] A. Li, W. Lai, and J. Kuang, "How do listeners identify creak? the effects of pitch range, prosodic position and creak locality in Mandarin," in *Proceedings of Speech Prosody 2022*, 2022, pp. 480–484.

[34] K. Dallaston and G. Docherty, "Estimating the prevalence of creaky voice: A fundamental frequency-based approach," in *Proceedings of the 19th International Congress of Phonetic Sciences*, S. Calhoun, P. Escudero, M. Tabain, and P. Warren, Eds. Australasian Speech Science and Technology Association Inc, 2019, pp. 581.1–5.

[35] T. Drugman, J. Kane, and C. Gobl, "Data-driven detection and analysis of the patterns of creaky voice," *Computer Speech and Language*, vol. 28, no. 5, pp. 1233–1253, 2014.

[36] O. Murton, S. Shattuck-Hufnagel, J.-Y. Choi, and D. D. Mehta, "Identifying a creak probability threshold for an irregular pitch period detection algorithm," *The Journal of the Acoustical Society of America*, vol. 145, no. 5, pp. EL379–EL385, 2019.

[37] J. Villegas, K. Markov, J. Perkins, and S. J. Lee, "Prediction of creaky speech by recurrent neural networks using psychoacoustic roughness," *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 2, pp. 355–366, 2020.

[38] 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.

Gender attitudes affect the strength of the Frequency Code

Sasha Calhoun, Paul Warren, Jemima Agnew, Joy Mills

Te Herenga Waka - Victoria University of Wellington (THW-VUW)

firstname.lastname@vuw.ac.nz

Abstract

Following the Frequency Code, certain intonational meanings have a biological basis: high vs. low pitch are physically linked to small vs. large body size and to female vs. male gender (via sexual dimorphism), leading to affective meanings like submissiveness vs. dominance. While such associations appear widespread, the code assumes culture- and individual-specific ideological links, e.g., between submissiveness and femininity. We present Implicit Association Test experiments measuring associations between voice pitch and body size/gender. All participants showed these associations, however, their strength varied according to listeners' genders and gender beliefs. We discuss implications for theories of pitch iconicity.

Index Terms: pitch perception, Frequency Code, sociophonetics, sound symbolism, iconicity, gender and language

1. Introduction

Since pitch is one of the few language features we share with other mammals and birds, it has been argued that an important source of intonational meaning is biological. In particular, the *Frequency Code* links high and low f0 with body size and sex, based on animal behaviour, and in turn to affective meanings such as submissiveness versus dominance [1, 2, 3]. Here we probe the Frequency Code from a sociophonetic perspective, where such meanings can be taken to be socially constructed. That is, they can stem from speakers' and listeners' experiences and beliefs in relation to users of different linguistic features, such as an ideological link between masculinity and dominance, which influences linguistic meanings related to low pitch (see [4, 5, 6]). Our proposal seeks to reconcile these approaches: the availability of iconic pitch associations depends not only on physical associations, but also on listeners' beliefs, particularly in relation to gender. We investigate this using Implicit Association Tests (IATs) [7], looking at associations between voice pitch and each of body size and gender, and how these vary by the gender and gender beliefs of the listener.

Biological codes have been proposed to "explain what is universal about the interpretation of pitch variation" ([3], p. 74), linking physiological properties associated with pitch production with informational and affective interpretations of pitch features in language. For example, larger animals have larger vocal apparatus, producing vocalisations with lower f0. Morton [1] observed mammals and birds tended to use lower f0 when acting aggressively, and higher f0 when acting submissively, regardless of their actual size, i.e. f0 conveyed *apparent* size. Following this, Ohala ([2], p. 327) proposes the Frequency Code: affective meanings such as "deference, politeness, submission, lack of confidence, are signaled by high and/or rising F0 whereas assertiveness, authority, aggression, confidence, threat are conveyed by low/or falling F0". Ohala [2] extended this to human sex differences, because of sexual dimorphism: since males tend to be larger than females, the relationship between low f0 and dominance is particularly associated with males, and high f0 and submissiveness with females. Importantly, while studies show that f0 may not be a reliable cue to human body size (e.g. [8, 9]), listeners strongly associate low pitch with larger body size [10]. Further, while gender and its performance are complex, and the links to biological sex indirect, listeners are very sensitive to vocal cues to gender in speech, including f0, from a very young age (e.g. [11, 12]). The Frequency Code has considerable support in perception studies, with high pitch associated with uncertainty, weakness, dependency and submissiveness across listeners of different cultural and linguistic backgrounds (e.g. [13, 14, 15, 16]). In particular, the link between low f0 and dominance is widely accepted in social psychology (e.g. [10, 17, 18]).

Biological code theory proposes that meanings of pitch features can stem from sound symbolism or iconicity, which may then be phonologized within a language [19, 2]. However, despite cross-linguistic commonalities in these seemingly natural links, the Frequency Code assumes culturally-specific ideological links, e.g. between submissiveness and femininity, which have largely not been accounted for in prior research. This runs contrary to decades of sociolinguistic and sociophonetic research showing that links between linguistic features and particular groups and affects are socially constructed (e.g. see [4, 5]). Nonetheless, recent work suggests how these approaches can be reconciled. For example, Eckert ([4], p. 754) discusses how iconic associations are "products of cultural constructions of resemblance to things in the natural world". Thus, of the multiple potential resemblances pitch (and other linguistic features) could have, those that align with existing listener 'world views' and the affordances of the context, are most likely to become established social meanings [5, 4, 20, 6]. For example, Walker et al. [20] note that the link between high pitch features and politeness depends on the link between high f0 and weakness or submissiveness, i.e. the speaker signals politeness by appearing weaker and non-threatening to their interlocutor. In a perception study on Korean and English, they show Korean listeners have similar pitch associations for submissiveness to English listeners, but not politeness, showing these are not linked in Korean; rather, politeness is usually signalled by low pitch features (see also [6]). Such an approach could also explain other apparently contradictory findings in Frequency Code research, e.g. dominance signalled by high pitch for female speakers [21].

Our research explores the idea that pitch iconicity can provide a shared 'extra-linguistic' basis for linguistic meanings. However, we propose that in language cognition the availability of different physical pitch associations, and thus different derived meanings, varies according to listeners' experiences and beliefs, and the context. In relation to the Frequency Code, the most important of these relate to listeners' gender and gender beliefs, as the core affective associations of the Frequency Code

(e.g. low pitch and dominance) align closely with normative gender. As a starting point, we look at associations of voice pitch with the core physical associations which are said to underlie the Frequency Code, i.e. body size and binary gender. We investigate whether the strengths of these associations differ between individuals based on their gender and gender beliefs. We predict that associations will be stronger for listeners of male gender and those with more normative, or less egalitarian, gender beliefs.

We investigate physical associations of voice pitch using IATs [7], a well-established task that measures implicit association strength between paired concepts and attributes. We measure associations of small/large body size or female/male with high/low pitch, using participants' speed at categorising stimuli in each category. Most previous studies reviewed above have used rating tasks. The IAT is arguably better to tap into subconscious associations which listeners may not have metacognition of [22]. The IAT has been shown to effectively measure subconscious associations of other linguistic features and, importantly, how these vary given social factors (e.g. [23, 24]).

2. Method

IAT experiments were run to find out firstly if participants showed the implicit associations between voice pitch and each of gender and body size predicted by the Frequency Code, and secondly, whether the strength of these associations were affected by the participants' gender and their experiences and beliefs around gender, as measured in gender attitude surveys.

2.1. Participants

Data is reported from 239 participants, recruited on Prolific (www.prolific.co), evenly distributed across experiment versions by participant gender. There were 113 female, 120 male and 6 non-binary/other gender, all with English as their first language and living in New Zealand, Australia or the United Kingdom. For the statistical modelling, female and non-binary/other genders were grouped (labelled 'non-male'). Median age was 33 years (range 18-69 years). Participants received Prolific credits for participation. This study was approved by the Human Ethics Committee at THW-VUW (#29710).

2.2. Materials

For the IAT experiments we created two sets of concept stimuli (gender and size) and one set of attribute stimuli (pitch). For gender classification, twelve names were chosen with strong associations to either female (*Anna, Claire, Jane, Julie, Rachel, Sarah*) or male (*Andrew, Daniel, David, James, Mark, Michael*) gender. Names were chosen using the New Zealand Department of Internal Affairs register of baby names (www.dia.govt.nz/diawebsite.nsf/wpg_URL/Services-Births-Deaths-and-Marriages-Most-Popular-Male-and-Female-First-Names). We selected names that were among the most frequent across the period 1960-2000, so they would be familiar to participants.

For size, twelve animals were chosen which would usually be classified as either small (*frog, mouse, blackbird, guinea pig, squirrel, rabbit*) or large (*elephant, giraffe, bear, crocodile, panda, camel*) in size. Black-and-white photographs of the animals were taken from the Animal Images database [25]. Pictures were all 300*225 pixels. The small and large sets were matched for familiarity and valence.

The pitch stimuli involved nonwords similar in phonetic make-up to discourse markers intended for later studies. The final selection for the current study was *ernerm* and *yerwer* (similar to *I mean* and *really*). Tokens of each nonword were recorded by three female and three male New Zealand English speakers in their thirties. The tokens were time- and amplitude-normalised, and then pitch-normalised using a purpose-built Matlab script to mean values of 195 Hz for females and 110 Hz for males (values based on means of word tokens in the New Zealand Spoken English Database [26]), with a declination slope of 0.7 ERB. High- and low-pitched versions were then created with all pitch values raised by 1.7 ERB or lowered by 0.7 ERB (values based on pre-testing by the authors).

All three sets of stimuli were chosen from larger sets after norming studies. The first (N=32) showed female names were rated 1.13 and male 6.78 on a scale 1 (definitely female) to 7 (definitely male). Participants' accuracy at categorising pitch stimuli (as high or low) was 97.9%. The second study (N=32) showed that small animal pictures were rated 1.41 and large 6.48 on a scale 1 (definitely small) to 7 (definitely large).

A questionnaire targeted participants' basic demographic information and language background. To gauge participants' gender attitudes and beliefs, questions were selected from established gender surveys. Participants had to indicate their level of agreement with statements using a Likert scale from 1 (strongly disagree) to 7 (strongly agree). We constructed five measures of gender attitudes: we selected the subsets of the Ambivalent Sexism Inventory [27] used in the New Zealand Attitudes and Values Survey (NZAVS) [28], i.e. five questions relating to Benevolent Sexism, e.g. "Women, compared to men, tend to have greater moral sensibility", and five relating to Hostile Sexism, e.g. "Women seek to gain power by getting control over men". Male Norms were measured with the five-item Male Norms Inventory [29], e.g. "Boys should prefer to play with trucks rather than dolls". Transgender/Non-Binary attitudes were measured via two statements adapted from the New Zealand Gender Attitudes Survey [30], e.g. "I would be comfortable with a transgender or non-binary person as a colleague", and finally, Social Dominance Orientation was measured using a six-item subset from the NZAVS inventory [31], e.g. "Inferior groups should stay in their place". This measure was included as voice pitch has been shown to relate to perceptions of social as well as physical dominance [10].

Table 1: Example sequence of blocks for IAT Experiments. Shows Consistent-First order

Block	No. of trials	Type	Items on left-key response	Items on right-key response
1	24	Practice	Male names	Female names
2	24	Practice	Low pitch	High pitch
3	4	Practice	Male + Low	Female + High
4	48	Consistent	Male + Low	Female + High
5	36	Practice	Female names	Male names
6	4	Practice	Female + Low	Male + High
7	48	Inconsistent	Female + Low	Male + High

2.3. Design and procedure

The IAT experiments were constructed and run in PsyToolkit, version 3.4 [32, 33], following a standard IAT design, as in [7, 34], see Table 1. In the practice blocks 1-2, participants learn to classify stimuli as being from each pair of a concept

and attribute, e.g. male/female and low/high pitch, linked to the left ('E') or right ('I') response key. In the third practice block and the following test block, concept and attribute stimuli are combined. In this example, these blocks are 'consistent', i.e. the expected pairing of concept and attribute, e.g. male and low pitch, are on the same response key. The response key for the concept is then reversed in Block 5, and then the reverse, 'inconsistent', combination of concept and attribute is tested (Blocks 6 and 7). If a participant has the expected implicit association between concept and attribute, they should be faster and more accurate at classifying stimuli in 'consistent' (4) than 'inconsistent' (7) blocks. Two concepts were included in each experiment version. The second concept followed the same design, except that the attribute practice block (2) is omitted as this has not changed. In blocks 1, 2, 4 and 7, each stimulus was repeated twice. In block 5 (reverse concept practice block), stimuli were repeated three times as [35] found this reduces order effects (see below). Four practice items were used for the combined blocks (3, 6). In all practice blocks stimulus order was randomised by participant. In test blocks, participants received one of nine pseudo-randomly ordered lists with no more than three responses with the same key response, or three concept or attribute stimuli, in a row.

Twenty-four versions of the IAT were constructed, involving three concepts, Gender, Size, and Effort (not reported here). All six possible pairings of the three concepts were used. Each version used only one voice gender: either the three male or the three female voices. Finally, both possible orderings of the consistent-inconsistent blocks were used, as it is well-established that the implicit association effect may be substantially smaller in inconsistent-consistent order (e.g. see [34, 35]). Table 1 shows consistent-inconsistent order, in inconsistent-consistent order Blocks 1, 3, 4 were reversed with 5-7.

Participants completed the experiment online using PsyToolkit, using a desktop or laptop in a quiet room with headphones. They completed the IAT first. The instructions and layout for the IAT closely followed those on the Project Implicit website (implicit.harvard.edu/implicit/takeatest.html). Participants first received instructions on the task and saw/heard all stimuli. Instructions were repeated before each block, although there was no break between the combined practice and test blocks. Each response had a timeout of 3s. If a participant responded incorrectly, they saw a red cross on the screen and then had to press the correct response. Participants then completed the demographic questions, followed by the gender attitude questions, presented in pseudo-random order. The experiment took approximately 25 minutes.

2.4. Analysis

The size of implicit associations for each participant for each concept-attribute pairing was gauged by D-scores, calculated with a script modified from [36] in R, based on [34]. Responses below 400ms were removed, and RTs for incorrect responses replaced with the participant's mean RT plus a 600ms 'penalty'. A D-score is the difference between a participant's mean response time in the inconsistent and consistent blocks (i.e. 7-4 in Table 1), divided by their SD in these blocks.

Linear regression models were built in R, with D-score as the dependent variable. Initial models included a five-way interaction of Concept (Gender or Size), Voice Gender (Male or Female voices in the experiment), Order (Consistent first or Inconsistent first), Half (first or second half of the experiment) and participant Gender (see further below). Elimination of non-significant effects used the *stepAIC* function in the MASS package with further manual elimination using *anova*, starting with higher-order interactions, until remaining factors and interactions were significant. Model estimates were extracted using *effects* and plotted with *ggplot2*.

Participant gender correlated with each of the gender attitude measures, which were also correlated with each other. Since correlated predictors can make regression models unreliable, we built separate models for participant gender and each of the gender attitude measures. Scores for each measure (Benevolent Sexism, Hostile Sexism, Male Norms, Transgender/Non-Binary Attitudes, Social Dominance Orientation) were derived by averaging ratings for the statements in each set, after reversing ratings for statements where a high rating indicated a more egalitarian attitude, so that higher indicated less egalitarian views for all statements. Median scores were: Benevolent Sexism 3.6, Hostile Sexism 2.6, Male Norms 2.2, Transgender/Non-Binary 1, Social Dominance 1.8.

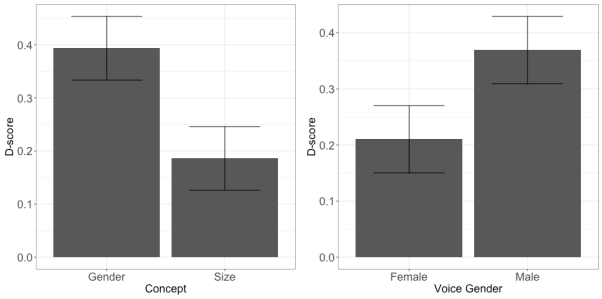


Figure 1: *Fitted D-scores by Concept (left) and Voice Gender (right). Error bars show standard error of the means.*

3. Results

3.1. Participant gender

The final regression model including Participant Gender showed simple effects of Concept ($F(1, 310) = 23.15$, $p<0.001$) and Voice Gender ($F(1, 310) = 12.25$, $p<0.001$), and interactions of Participant Gender*Order ($F(1, 310) = 9.15$, $p=0.003$) and Half*Order ($F(1, 310) = 5.71$, $p=0.017$). As can be seen in Figure 1, D-scores were higher, indicating a stronger effect, for the Gender than Size concept ($t=-4.81$, $p<0.001$), and for Male than Female Voices ($t=6.35$, $p<0.001$).

The Participant Gender*Order interaction can be seen in Figure 2 (left). Comparisons using *emmeans* (fdr method) showed higher D-scores in Consistent-first order than Inconsistent-first for Male ($t=5.96$, $p<0.001$) but not Non-male participants ($p=0.11$). Further, Male participants had higher D-scores than Non-male in Consistent-first order ($t=-3.81$, $p<0.001$) but not Inconsistent-first order ($p=0.69$). As noted above, D-scores are often found to be higher in Consistent-first than Inconsistent-first order, although the interaction with gender was not predicted (see further below). For the Half*Order interaction, comparisons showed D-scores were higher in Consistent-first order than Inconsistent-first in both the first ($t=5.5$, $p<0.001$) and second half ($t=2.14$, $p=0.039$) of the experiment. D-scores were also higher in the first half than second in Consistent-first order ($t=2.43$, $p=0.023$) but not Inconsistent-first ($p=0.36$). This is consistent with a learning effect, where D-scores were smaller in the second half of the experiment as participants got used to the task, although this could only be seen in Consistent-first order.

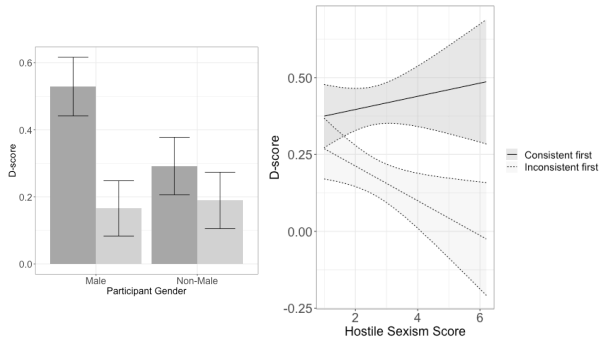


Figure 2: *Fitted D-scores by Participant Gender and Order (left) and Hostile Sexism score and Order (right). Error bars show standard error of the means.*

3.2. Gender attitudes measures

As explained in section 2.4, separate models were built for each of the five gender attitude measures. The models for Benevolent Sexism, Hostile Sexism, Male Norms and Social Dominance Orientation returned simple effects of Concept, Voice Gender and an interaction of Half*Order in line with the Participant Gender model above (details not reported for space reasons). The Hostile Sexism model also included an interaction of Hostile Sexism and Order ($F(1, 310) = 4.59$, $p=0.033$). As seen in Figure 2 (right), participants with low Hostile Sexism scores show a small difference between D-scores in Consistent-first and Inconsistent-first orders. As Hostile Sexism scores increase, D-scores in Consistent-first order rise while those in Inconsistent-first order fall. This was not predicted, but is in line with the Participant Gender*Order interaction above.

For Social Dominance Orientation, the interaction with Order was marginal ($F(1, 310) = 3.6$, $p=0.059$) and similar to that for Hostile Sexism. The Benevolent Sexism and Male Norms models returned only simple effects of the measure ($F(1, 311) = 12.89$, $p<0.001$ and $F(1, 311) = 7.79$, $p=0.006$, respectively). As measure scores rise, D-scores fall. The lack of interactions with Order is a likely result of loss of power, since each gender measure was modelled separately. As seen in Figure 2, the reduction of D-scores with increasing Hostile Sexism scores in Inconsistent-first order was larger than the rise in Consistent-first order, so this amounts to an overall reduction. The Transgender/Non-Binary Attitudes model returned a Half*Order interaction and a four-way interaction of Concept*Voice Gender*Transgender/Non-Binary Attitudes*Order ($F(1, 300) = 4.01$, $p=0.046$). While there is not space to explore the four-way interaction, broadly the same pattern of Gender Measure*Order was found as in Figure 2, but not for all Concept*Voice Gender combinations.

4. Discussion

This study used IATs to explore implicit associations of voice pitch predicted by the Frequency Code. As predicted, low pitch is implicitly associated with males and large body size, and high pitch with females and small body size. Further findings were generally consistent with our proposal that these associations would be affected by participants' experiences and beliefs, particularly in relation to gender. The implicit association with pitch was stronger for gender than body size, and for male voices than female. Arguably, cultural stereotypes informing these associations are stronger for gender and male voices than size and female voices. For participant gender and most gender attitude measures, we found an unexpected interaction with IAT block order. When the blocks with the 'consistent' associations (e.g. low pitch and male) were presented first, we found the predicted effect: implicit associations were stronger for males and those with less egalitarian gender attitudes. However, in inconsistent-first order, a different pattern emerged: non-males and those with more egalitarian gender attitudes showed no or a very small difference to consistent-first order, while males and those with less egalitarian attitudes showed a much smaller effect. We discuss each of these findings further below.

Implicit associations of body size and gender with voice pitch are based on physical associations. However the association with body size is arguably primary, existing in animal communication systems even when species do not show sexual dimorphism [1]. Nonetheless, we found stronger implicit associations for gender than body size. We submit this fits with our general proposal: cultural stereotypes relating to gender are stronger and more visible than those involving body size (in connection with voice pitch). Similarly, implicit associations for male voices were stronger than female. We suggest this is because, as the historically dominant and privileged gender, ideologies relating to the Frequency Code are more entrenched for males, so the physical associations on which they are based are more salient. For females, given the impact of feminism, these are more challenged, and the iconic links are disrupted and weaker. This matches previous findings that Frequency Code associations do not always hold for female speakers [9, 21, 18].

As predicted, we found stronger implicit associations between voice pitch and body size/gender for males and those with less egalitarian gender beliefs. We suggest this is because the iconic pitch associations based on the Frequency Code largely align with normative gender ideology, e.g. associations between masculinity and dominance. Therefore physical associations on which the iconic associations rest will be stronger for individuals with normative beliefs about gender. However, unexpectedly, we found this effect only when the IAT task was presented in 'consistent-first' order, i.e. with blocks showing the expected voice pitch associations first. We believe that this result does make sense given our proposal. For listeners with likely stronger a priori biases, i.e., males and those with less egalitarian beliefs, experiencing the consistent associations first (male names with low pitch, female with high) reinforces or at least does not contradict these connections. If, however, they experience the inconsistent associations in the first block, this contradicts their a priori bias and leads to a reduction in its effect on the task. These effects persists through the experiment.

Our results show IATs, combined with measures of individual differences, are a promising way to investigate how iconic associations of voice pitch may form part of language cognition and linguistic meaning. In future work, we plan to investigate implicit associations of linguistic features involving pitch, such as uptalk and creaky voice, which have been argued to have Frequency Code-related meanings (see [37]). We will also explore how to more effectively quantify the effect of gender and gender attitude measures in our modelling, e.g. principal components analysis, rather than treating these separately.

We believe that our proposal has the potential to contribute significantly to the understanding of intonational meaning and how this is affected by biological, social, and individual (experiential) factors, as well as to longstanding debates on the (non-)arbitrariness of linguistic meaning. By quantifying how voice pitch can be simultaneously 'natural' and 'social', we offer a new approach to investigating iconicity in speech.

5. Acknowledgements

This work was funded by a Faculty Strategic Research Grant from THW-VUW. Thanks to Jen Hay, Abby Walker, Joe Bulbulia and Gina Grimshaw for advice and to the participants.

6. References

[1] E. S. Morton, "On the occurrence and significance of motivation-structural rules in some bird and mammal sounds," *The American Naturalist*, vol. 111, no. 981, pp. 855–869, 1977.

[2] J. J. Ohala, "The frequency code underlies the sound-symbolic use of voice pitch," in *Sound Symbolism*, L. Hinton, J. Nichols, and J. J. Ohala, Eds. Cambridge, UK: Cambridge University Press, 1994, pp. 325–347.

[3] C. Gussenhoven, "Paralinguistics: Three biological codes," in *The Phonology of Tone and Intonation*. Cambridge: Cambridge University Press, 2004, pp. 71–96.

[4] P. Eckert, "The limits of meaning: Social indexicality, variation, and the cline of interiority," *Language*, vol. 95, no. 4, pp. 751–776, 2019.

[5] A. D'Onofrio and P. Eckert, "Affect and iconicity in phonological variation," *Language in Society*, vol. 50, no. 1, pp. 29–51, 2021.

[6] B. Winter, G. E. Oh, I. Hbscher, K. Idemaru, L. Brown, P. Prieto, and S. Grawunder, "Rethinking the frequency code: a meta-analytic review of the role of acoustic body size in communicative phenomena," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 376, no. 1840, p. 20200400, 2021.

[7] A. G. Greenwald, D. E. McGhee, and J. L. K. Schwartz, "Measuring individual differences in implicit cognition: The Implicit Association Test," *Journal of Personality and Social Psychology*, vol. 74, no. 6, pp. 1464–1480, 1998.

[8] D. Rendall, S. Kollias, C. Ney, and P. Lloyd, "Pitch (F0) and formant profiles of human vowels and vowel-like baboon grunts: The role of vocalizer body size and voice-acoustic allometry," *The Journal of the Acoustical Society of America*, vol. 117, no. 2, pp. 944–955, 2005.

[9] W. A. van Dommelen and B. H. Moxness, "Acoustic parameters in speaker height and weight identification: Sex-specific behaviour," *Language and Speech*, vol. 38, no. 3, pp. 267–287, Jul. 1995.

[10] M. Armstrong, A. Lee, and D. Feinberg, "A house of cards: bias in perception of body size mediates the relationship between voice pitch and perceptions of dominance," *Animal Behaviour*, vol. 179, pp. 43–51, 2019.

[11] S. B. Most, A. V. Sorber, and J. G. Cunningham, "Auditory Stroop reveals implicit gender associations in adults and children," *Journal of Experimental Social Psychology*, vol. 43, no. 2, pp. 287–294, 2007.

[12] L. Nagels, E. Gaudrain, D. Vickers, P. Hendriks, and D. Bakent, "Development of voice perception is dissociated across gender cues in school-age children," *Scientific Reports*, vol. 10, no. 1, p. 5074, 2020.

[13] B. Borkowska and B. Pawlowski, "Female voice frequency in the context of dominance and attractiveness perception," *Animal Behaviour*, vol. 82, no. 1, pp. 55–59, 2011.

[14] W. Gu, P. Tang, K. Hirose, and V. Auberg, "Crosslinguistic comparison on the perception of Mandarin attitudinal speech," in *Interspeech 2015*. ISCA, 2015, pp. 1334–1338.

[15] C. Gussenhoven and A. Chen, "Universal and language-specific effects in the perception of question intonation," in *Proc. of ICSLP 2000*, Beijing, 2000, pp. 91–94.

[16] R. van Bezooijen, "Sociocultural aspects of pitch differences between Japanese and Dutch women," *Language and Speech*, vol. 38, pp. 253–265, 1995.

[17] K. Pisanski and G. Bryant, "The evolution of voice perception," in *The Oxford Handbook of Voice Studies*, N. S. Eidsheim and K. Meizel, Eds. Oxford: Oxford University Press, 2019.

[18] M. S. Tsantani, P. Belin, H. M. Paterson, and P. McAleer, "Low vocal pitch preference drives first impressions irrespective of context in male voices but not in female voices," *Perception*, vol. 45, no. 8, pp. 946–963, 2016.

[19] M. Dingemanse, D. E. Blasi, G. Lupyan, M. H. Christiansen, and P. Monaghan, "Arbitrariness, iconicity, and systematicity in language," *Trends in Cognitive Sciences*, vol. 19, no. 10, pp. 603–615, 2015.

[20] A. Walker, J. J. Holliday, M. Jung, and E. Cho, "A closer look at the sound of politeness in Korean," in *New Ways of Analyzing Variation Asia Pacific (NWAV-AP) 6*, Singapore, 2021.

[21] P. McAleer, A. Todorov, and P. Belin, "How do you say 'Hello'? Personality impressions from brief novel voices," *PLOS ONE*, vol. 9, no. 3, p. e90779, 2014.

[22] B. Kurdi, K. A. Ratliff, and W. A. Cunningham, "Can the Implicit Association Test serve as a valid measure of automatic cognition? A response to Schimmack (2021)," *Perspectives on Psychological Science*, vol. 16, no. 2, pp. 422–434, Mar. 2021.

[23] K. Campbell-Kibler, "The Implicit Association Test and sociolinguistic meaning," *Lingua*, vol. 122, no. 7, pp. 753–763, 2012.

[24] P. lvarez Mosquera, "The use of the Implicit Association Test (IAT) for sociolinguistic purposes in South Africa," *Language Matters*, vol. 48, no. 2, pp. 69–90, 2017.

[25] C. Possidnio, J. Graa, J. Piazza, and M. Prada, "Animal Images Database: Validation of 120 images for human-animal studies," *Animals*, vol. 9, no. 8, p. 475, 2019.

[26] P. Warren, "NZSED: building and using a speech database for New Zealand English," *New Zealand English Journal*, vol. 16, pp. 53–58, 2002.

[27] P. Glick and S. T. Fiske, "The Ambivalent Sexism Inventory: Differentiating hostile and benevolent sexism," *Journal of Personality and Social Psychology*, vol. 70, no. 3, pp. 491–512, 1996.

[28] C. G. Sibley, "Sampling procedure and sample details for the New Zealand Attitudes and Values Study," New Zealand Attitudes and Values Study, Tech. Rep., 2021. [Online]. Available: http://nzavs.auckland.ac.nz

[29] R. C. McDermott, R. F. Levant, J. H. Hammer, N. C. Borgogna, and D. K. McKelvey, "Development and validation of a five-item Male Role Norms Inventory using bifactor modeling," *Psychology of Men & Masculinities*, vol. 20, no. 4, pp. 467–477, 2019.

[30] National Council of Women of NZ, "New Zealand Gender Attitudes Survey," 2019. [Online]. Available: https://genderequal.nz/ga-survey/

[31] J. Sidanius and F. Pratto, *Social Dominance: An Intergroup Theory of Social Hierarchy and Oppression*. Cambridge: Cambridge University Press, 1999.

[32] G. Stoet, "PsyToolkit: A software package for programming psychological experiments using Linux," *Behavior Research Methods*, vol. 42, no. 4, pp. 1096–104, Nov. 2010.

[33] ——, "PsyToolkit: A novel web-based method for running online questionnaires and reaction-time experiments," *Teaching of Psychology*, vol. 44, no. 1, pp. 24–31, 2017.

[34] A. G. Greenwald, B. A. Nosek, and M. R. Banaji, "Understanding and using the Implicit Association Test: I. An improved scoring algorithm." *Journal of Personality and Social Psychology*, vol. 85, no. 2, pp. 197–216, 2003.

[35] B. A. Nosek, A. G. Greenwald, and M. R. Banaji, "Understanding and Using the Implicit Association Test: II. Method Variables and Construct Validity," *Personality and Social Psychology Bulletin*, vol. 31, no. 2, pp. 166–180, 2005.

[36] J. Röhner and P. J. Thoss, "A tutorial on how to compute traditional IAT effects with {R}," *The Quantitative Methods for Psychology*, vol. 15, no. 2, pp. 134–147, 2019.

[37] H. White and S. Calhoun, "Mediated iconicity: Effect of age on affective associations of uptalk and creak," in *18th Conference on Laboratory Phonology*, 2022.

Is there an Influence of Autistic Traits on Audio-Visual Speech & Song Emotion Perception?

Evie Day, Jia Hoong Ong

University of Reading, Reading, United Kingdom

e.r.day@student.reading.ac.uk, jiahoong.ong@reading.ac.uk

Abstract

Autistic individuals are reported to have atypical emotion perception. Evidence indicates this may be domain specific. However, most previous studies have used unimodal stimuli and have not explored perception of emotions in sung vocals. This study thus examined whether emotion recognition differs as a function of autistic traits using audio-visual speech and song stimuli. We found no general impairment in emotion perception across domains as a function of autistic traits, though higher autistic traits were associated with slower reaction time for some emotions. Findings are interpreted in light of methodological differences with previous studies, particularly our use of audio-visual stimuli.

Index Terms: Autism, emotion, music, speech, song, audio-visual

1. Introduction

Autism spectrum condition is a neurodevelopmental condition characterised by differences in social communication and the presence of restricted, repetitive, and stereotyped behaviours [1]. Though not a formal diagnostic criterion, various studies have indicated that autistic individuals find emotion processing more challenging than neurotypical (NT) individuals, which could exacerbate the challenges autistic individuals face in social communication [2]. (In this paper, we respectfully use the term 'autistic individuals', a term that seems to be preferred by most individuals with autism [3].)

Autistic individuals appear to find emotion perception challenging across most modalities and stimuli. Poorer performance in emotion recognition among autistic individuals or individuals with higher levels of autistic traits has been repeatedly demonstrated with static photographs and dynamic video-only clips of human faces as stimuli (see [4] for a meta-analysis on this topic). Curiously, group differences seem to disappear when recognising emotions in nonhuman/cartoon-like faces [5], suggesting that performance could be modulated by how socially meaningful the stimuli are. Similar to findings with human faces as stimuli, autistic individuals tend to show atypical perception of emotions in speech, when compared to NT individuals (see [6] for a review). While this could be due to language difficulties in some autistic individuals [7], even when matched to their NT counterparts on verbal ability, autistic individuals seem to have found decoding of emotions in speech more challenging [8]. Given that emotions in speech are mostly conveyed through acoustic features such as pitch, amplitude, and tempo/duration [9], some suggest that speech emotion processing ability may be related to one's basic auditory perception which has been found to be atypical among autistic individuals, particularly for vocal stimuli [10], [11].

Considerably fewer studies have investigated whether autistic and NT individuals differ in their processing of emotions in music. The tentative conclusion is that there are no group differences [12], suggesting music may be processed differently to speech. A recent meta-analysis found that difficulties in emotion processing in music among autistic individuals may apply to fear and sadness, but not happiness [13]. However, this should be interpreted with caution given the small number of studies with musical stimuli included in the meta-analysis [13]. Preserved music-emotion processing and impaired speech-emotion processing among autistic individuals is not surprising, given the double dissociation in processing the two domains among autistic individuals. For example, autistic individuals seem to have equal, if not superior, processing of musical contour and intervals, and pitch memory [14][15]. Indeed, the apparent music processing ability among autistic individuals has led some to suggest music-based interventions for autism [16]. If music emotion perception is associated with basic auditory perception, as with speech emotion perception, then it follows that autistic individuals may show a preserved ability to decode emotions in music.

It should be noted that most of the music-related studies reviewed above used stimuli generated from musical instruments, not including sung vocals. Drawing parallels to differences in human vs. non-human faces, it is unclear if findings from music emotion processing generated from musical instruments would generalise to that of human-produced sung vocals, the latter of which may be more socially meaningful. Moreover, most studies on emotion perception among autistic individuals have relied on unimodal stimuli (e.g., static photographs, audio-only clips, etc.), which are arguably less ecologically valid than the use of audio-visual (AV) stimuli. Of the handful of studies that examined AV emotion processing among autistic and NT individuals, no groups differences in accuracy were reported [17], [18], though autistic individuals showed less gain from multimodal relative to unimodal stimuli than NT individuals [18].

When considering the effect of autism or autistic traits on emotion processing, it is vital to account for alexithymia, a condition related to difficulty describing and identifying emotions, especially since alexithymia is highly prevalent among autistic individuals [19]. Indeed, some argue that it is alexithymia, rather than autism per se, that drives the group differences in emotion perception seen in previous studies [20]. Thus, alexithymia needs to be accounted for, to determine whether autism or autistic traits may have any influence on emotion processing above and beyond alexithymia.

The current study addresses the gaps identified in the literature: using AV stimuli, we examined emotion perception in speech and song with adults with varying levels of autistic traits, while controlling for their alexithymia. Based on previous studies, we hypothesised a negative effect of autistic

traits on speech emotion perception. We made no clear predictions about the effect of autistic traits on song emotion perception: if performance on the song condition is based on music auditory perception, then no effect of autistic traits is expected; however, if performance is based on the social nature of the song stimuli, then a negative effect of autistic traits is expected.

2. Methods

2.1. Participants

Participants consisted of 55 young adults (M_{age} = 24.00, SD_{age} = 4.04, Range = 18-30; Female n = 34, Male n = 20, Non-binary n = 1), all of whom self-reported to have normal or corrected-to-normal vision and hearing. Most were native British English speakers (n = 49) and of the few who were not, they rated their English proficiency to be above-average (i.e., rating themselves at least a 6 on a 7-point scale).

Approximately half of the participants self-reported to have a formal diagnosis of autism (n = 27), though this was not verified due to the anonymity of the online experiment. Somewhat confirming their diagnosis, the autistic participants had a significantly higher autistic traits score (as measured using the Autism Spectrum Quotient, or AQ – see Tasks & Stimuli subsection) than the neurotypical participants ($t(53)$ = 6.56, p < .001). Some participants (n = 15) reported having experience with musical training: their mean cumulative experience ranged from 0.5 to 35 years (M = 9.37, SD = 8.61) and some participants (n = 9) reported having experience with drama and acting, with their cumulative experience ranging between 1 and 14 years (M = 5.89, SD = 5.18).

Participants were recruited from the Psychology research participant pool or from Prolific. All participants provided their written informed consent prior to their participation, and they were given course credit or monetary compensation as reimbursement for their time. The study protocol was reviewed and approved by the University Research Ethics Committee at the University of Reading.

2.2. Tasks and stimuli

2.2.1. Autism Spectrum Quotient (AQ)

Autistic traits were measured using The Autism Spectrum Quotient (AQ) [21], a screening test designed to determine the degree to which adults have traits associated with autism such as social skills, attention to detail, and communication. The AQ consists of 50-items, for which participants are required to indicate how much they agree or disagree on a 4-point scale. Higher AQ scores reflect higher levels of autistic traits.

2.2.2. Toronto Alexithymia Scale (TAS)

Alexithymia was measured using the Toronto Alexithymia Scale (TAS) [22]. This 20-item self-report instrument measures difficulty identifying and describing emotions by asking participants to how much they endorse each item on a 5-point scale. Higher TAS scores suggest more alexithymic traits.

2.2.3. Emotion recognition task

Stimuli from the emotion recognition task were taken from the Reading Everyday Emotion Database (REED) [23]. The REED consisted of AV recordings of native British English adults with varying levels of acting experience ('encoders') expressing emotions using their 'everyday' recording devices (e.g., webcam, mobile phone). A subset of 160 recordings from the REED was used in the task: 2 domains (speech/song) × 10 emotions (1 neutral; 6 basic—angry, happy, sad, disgusted, fearful, surprised; and 3 complex— embarrassed, sarcastic, and stressed) × 8 encoders. All the stimuli have the same verbal content: "Happy birthday to you" that is either spoken or sung to the first line of the Happy Birthday song. The recordings were recognised above chance in a previous validation study.

On every trial, participants were presented with a fixation cross for 500ms, then the recording, after which they were asked to select what emotion they thought was being portrayed from a choice of six labels (the target, 4 foils, and an 'Other' option, which should be chosen should neither of the 5 labels adequately describe the depicted emotion). Participants could not replay the recording, and they had to respond within 8s; otherwise, an incorrect response was recorded, and the next trial began automatically. Participants were given the opportunity to take a break after every 20 clips. Participants completed four practice trials prior to the start of the task.

To ensure participants were paying attention, catch trials were inserted throughout the task, consisting of a greyscale clip with no audio or a clip with an auditory beep. Participants were instructed to select 'Other' for those catch trials.

2.3. Procedure

Participants completed the tasks online, hosted on Gorilla [24]. After providing their informed consent and checking that the volume was at a comfortable level, participants completed a short demographic questionnaire, followed by the AQ and TAS questionnaires. Finally, participants completed the emotion recognition task. The entire study took approximately 30 minutes to complete.

2.4. Data analysis

Analysis was conducted on accuracy and reaction time on the emotion recognition task, separately.

Accuracy data was modelled using a binomial mixed effects model, with the binary variable Correct (Correct/Incorrect) as the dependent variable. We entered Domain (Speech/Song), Emotion (all 10 emotions), autistic traits (AQ), and all the possible interactions between them. To account for the possible effects of alexithymia, we included their TAS score as a predictor too. As random effects, we entered by-item random intercepts, by-subject random intercepts, and by-subject random slopes for Domain and Emotion.

Reaction time (RT) data on correct trials was modelled using a linear mixed effects model with the dependent variable Reaction Time in milliseconds. The same fixed effects as the Accuracy model were entered. As random effects, only by-subject and by-item random intercepts were entered due to convergence issues.

In both models, all categorical variables were effect coded and all continuous variables were mean centred. We analysed both models using the *lme4* package [25]. Statistical significance of each predictor was determined using the *Anova()* function from the *car* package [26]. Subsequent pairwise comparisons were conducted using the *emmeans* package with Tukey correction [27]. Model reduction for nonsignificant predictors was not done as the predictors were related to our a priori predictions.

3. Results

3.1. Accuracy

Output of the accuracy model is displayed in Table 1.

Table 1. *Output of the accuracy model.*

Predictors	χ^2	df	*p*
Intercept	0.02	1	.883
Domain	1.44	1	.230
Emotion	26.34	9	.002
AQ	0.11	1	.738
TAS	0.43	1	.513
Domain × AQ	0.70	1	.403
Domain × Emotion	10.96	9	.278
Emotion × AQ	9.42	9	.399
Domain × Emotion × AQ	15.34	9	.082

There was a main effect of Emotion, such that, as shown in Figure 1, some of the basic emotions (Happy, Neutral) were recognised more accurately than some of the complex emotions (Embarrassed, Stressed; Happy vs. Embarrassed: $z = 3.97$, $p = .003$; Happy vs. Stressed: $z = 3.45$, $p = .020$; Neutral vs. Embarrassed: $z = 3.59$, $p = .012$).

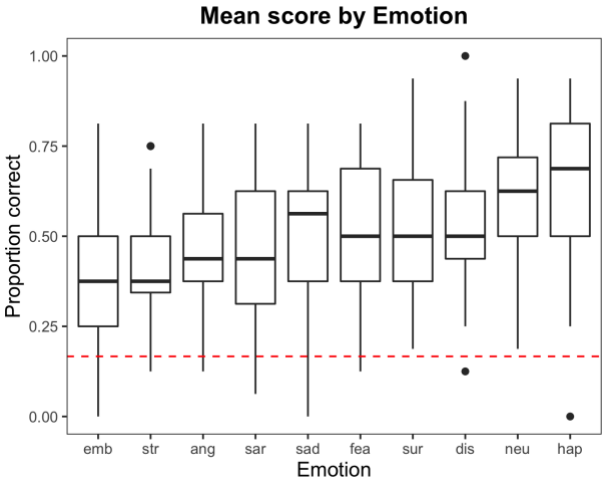


Figure 1: *Boxplots depicting accuracy by emotion. Dashed line represents chance level (1/6).*

We also explored the marginal three-way interaction between Domain, Emotion, and AQ, and we found that Surprised Song condition had a marginally significant negative AQ slope, but Disgusted Song condition had a marginally positive AQ slope (Disgusted: $z = 1.93$, $p = .054$; Surprised: $z = 1.95$, $p = .051$). Pairwise comparisons across domains only revealed a significant difference in the Disgusted conditions ($z = 2.243$, $p = .025$) but not in the Surprised conditions ($z = 1.56$, $p = .118$, see Figure 2).

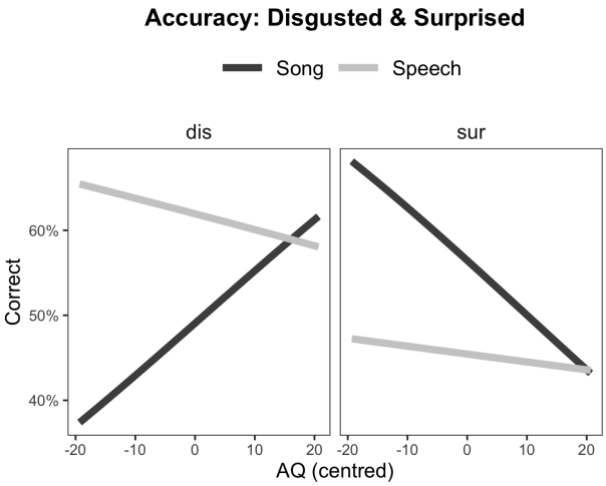


Figure 2: *Predicted percent correct for Disgusted (left) and Surprised (right) Speech (light grey) and Song (dark grey) as a function of autistic traits (AQ).*

3.2. Reaction time

Table 2 displays the output of the RT model. There was a marginal effect of TAS such that, surprisingly, higher TAS scores led to faster RT ($B = -13.99$, SE = 7.41).

Table 2. *Output of the RT model.*

Predictors	χ^2	df	*p*
Intercept	1273.52	1	<.001
Domain	0.15	1	.697
Emotion	33.04	9	<.001
AQ	1.93	1	.165
TAS	3.56	1	.059
Domain × AQ	2.72	1	.099
Domain × Emotion	14.04	9	.121
Emotion × AQ	20.75	9	.014
Domain × Emotion × AQ	8.76	9	.460

There was a main effect of Emotion, but this was qualified by an Emotion × AQ interaction. Follow-up comparisons revealed that the effect of AQ was only significant for Sarcastic ($z = 2.38$, $p = .017$) and Surprised ($z = 2.66$, $p = .008$) such that higher AQ was associated with longer RT for those two emotions (see Figure 3).

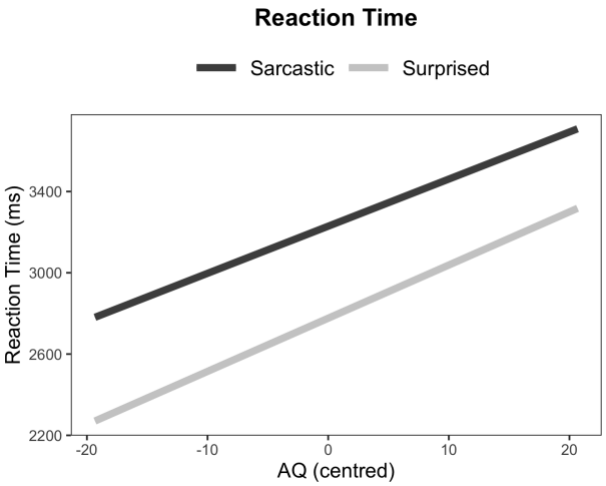


Figure 3: *Predicted RT (in ms) for Sarcastic (dark grey) and Surprised (light grey) as a function of autistic traits (AQ).*

4. Discussion

This study used AV stimuli to examine whether emotion perception in speech and song may differ as a function of autistic traits (AQ), while controlling for alexithymia. We predicted that there would be a negative AQ relationship for speech but made no clear predictions for emotion perception in song. On both accuracy and reaction time (RT), we found no general effect of AQ. A closer inspection revealed some weak evidence that the effect of AQ on accuracy might be modulated by Domain and Emotion. Results also suggested that AQ effect on RT was modulated by Emotion.

Unpacking the marginal interaction between AQ, Domain and Emotion for accuracy revealed that individuals with high AQ showed comparable performance for Speech and Song for Disgusted and Surprised, whereas individuals with low AQ showed differential performance: better for Speech than for Song for Disgusted, and the reverse for Surprised (see Figure 2). It is not immediately clear why this is the case, but we caution against drawing too strong a conclusion from this interaction given that the interaction and the subsequent post-hoc comparisons were only marginally significant.

In terms of the AQ × Emotion interaction for RT, we found that higher AQ was associated with longer RT for Sarcastic and Surprised. These findings are particularly interesting in relation to generally poorer accuracy performance for Surprised among individuals with high AQ. Taken together, these results suggest that Surprised, and to an extent, Sarcastic, may be challenging for individuals with higher levels of autistic traits, consistent with previous studies [28], [29]. These emotions to an extent rely on Theory of Mind, or the ability to understand one's mental states to infer their feelings or thoughts, which is thought to be atypical among autistic individuals [30]. For example, to recognise that someone is surprised, one must infer that they are unaware of or did not expect the situation that is unfolding.

We did not observe a general effect of AQ on emotion recognition of speech and song. On the latter, this is in line with the idea that music emotion processing among autistic individuals is spared [12]. This is to be expected if music emotion processing relies in part on one's basic auditory processing ability, which is found to be preserved among autistic individuals for musical stimuli [14], [15]. The lack of a negative effect of autistic traits on speech emotion recognition, on the other hand, is surprising, given previous studies reporting group differences when decoding affective speech prosody [6]. We propose that this may be due to methodological differences: previous studies have mostly relied on unimodal stimuli (e.g., audio-only speech affective prosody) whereas we used audio-visual stimuli. The AV benefit is well-documented, at least among neurotypicals [31]. Suppose that speech emotion perception is associated with autistic individuals' atypical basic auditory vocal perception, as suggested by previous studies [10], [11]. Then, our findings, together with the few AV studies on autism [17], [18], suggest that any emotion processing difficulties that autistic individuals or individuals with high autistic traits may have may be ameliorated with the presence of (redundant) information from the visual modality. This remains speculative at this stage of course, and further studies are needed to confirm this (indeed, we plan to repeat the study with an audio-only condition using the same stimulus set).

Some limitations of the study are worth noting. The sample size is modest, and given the number of predictors in the model, a larger sample size may be necessary to have sufficient power to detect any effects. Moreover, we used a dimensional (autistic traits) approach rather than the typical case-control approach seen in previous autism studies, comparing individuals who have received an autism diagnosis with those who did not. Indeed, some have cautioned against conflating autistic traits obtained from self-report measures with autism [32]. When we repeated the analysis with self-reported autism diagnosis, that is, we compared those with vs. without autism (though note that we were unable to confirm their diagnosis), we found no effects or interactions involving diagnosis. Thus, the case-control approach similarly revealed no group differences in emotion processing of speech and song when audio-visual stimuli were used. Future studies should also consider how gender may influence the findings, given some evidence for gender differences in emotion processing [33]. This was not possible in the present study given the imbalance of gender distribution (just over a third of our sample identified as male, one non-binary individual, and the rest female). With more participants, we hope to investigate this further by including biological sex and gender in the analysis to examine how they may differ in emotion processing, and how this might relate to autism.

5. Conclusions

In conclusion, contrary to our prediction and previous studies, we did not find any evidence of differences in general emotion perception in speech and song as a function of autistic traits, though individuals with high autistic traits were slower at recognising emotions that require Theory of Mind. We speculate that the lack of a general effect of autistic traits relative to previous studies may be due to methodological differences: specifically, unlike previous studies that have mostly used unimodal stimuli, our use of audio-visual stimuli may have ameliorated any differences in emotion processing. Further work is needed to confirm this, which, if true, will have implications on whether the findings of emotion processing differences among autistic individuals in previous studies reflect a true atypical ability or an artefact of stimulus modality.

6. Acknowledgements

This work is supported by the European Union's Horizon 2020 research and innovation programme under the Marie

Skłodowska-Curie grant agreement No. 887283 awarded to JHO.

7. References

[1] American Psychiatric Association, 'Diagnostic and statistical manual of mental disorders (5th ed.)'. 2013. [Online]. Available: https://doi.org/10.1176/appi.books.9780890425596

[2] B. T. Williams and K. M. Gray, 'The relationship between emotion recognition ability and social skills in young children with autism', *Autism*, vol. 17, no. 6, pp. 762–768, Nov. 2013, doi: 10.1177/1362361312465355.

[3] L. Kenny, C. Hattersley, B. Molins, C. Buckley, C. Povey, and E. Pellicano, 'Which terms should be used to describe autism? Perspectives from the UK autism community', *Autism*, vol. 20, no. 4, pp. 442–462, May 2016, doi: 10.1177/1362361315588200.

[4] M. Uljarevic and A. Hamilton, 'Recognition of emotions in autism: A formal meta-analysis', *J. Autism Dev. Disord.*, vol. 43, no. 7, pp. 1517–1526, 2013, doi: 10.1007/s10803-012-1695-5.

[5] D. B. Rosset, C. Rondan, D. Da Fonseca, A. Santos, B. Assouline, and C. Deruelle, 'Typical Emotion Processing for Cartoon but not for Real Faces in Children with Autistic Spectrum Disorders', *J. Autism Dev. Disord.*, vol. 38, no. 5, pp. 919–925, May 2008, doi: 10.1007/s10803-007-0465-2.

[6] M. Zhang, S. Xu, Y. Chen, Y. Lin, H. Ding, and Y. Zhang, 'Recognition of affective prosody in autism spectrum conditions: A systematic review and meta-analysis', *Autism*, p. 1362361321995725, Mar. 2021, doi: 10.1177/1362361321995725.

[7] T. May *et al.*, 'Trends in the Overlap of Autism Spectrum Disorder and Attention Deficit Hyperactivity Disorder: Prevalence, Clinical Management, Language and Genetics', *Curr. Dev. Disord. Rep.*, vol. 5, no. 1, pp. 49–57, Mar. 2018, doi: 10.1007/s40474-018-0131-8.

[8] S. Fridenson-Hayo *et al.*, 'Basic and complex emotion recognition in children with autism: Cross-cultural findings', *Mol. Autism*, vol. 7, no. 52, pp. 1–11, 2016, doi: 10.1186/s13229-016-0113-9.

[9] P. N. Juslin and P. Laukka, 'Communication of emotions in vocal expression and music performance: Different channels, same code?', *Psychol. Bull.*, vol. 129, no. 5, pp. 770–814, 2003, doi: 10.1037/0033-2909.129.5.770.

[10] S. Schelinski and K. von Kriegstein, 'The relation between vocal pitch and vocal emotion recognition abilities in people with autism spectrum disorder and typical development', *J. Autism Dev. Disord.*, vol. 49, no. 1, pp. 68–82, Jan. 2019, doi: 10.1007/s10803-018-3681-z.

[11] K. O'Connor, 'Auditory processing in autism spectrum disorder: A review', *Neurosci. Biobehav. Rev.*, vol. 36, no. 2, pp. 836–854, 2012, doi: 10.1016/j.neubiorev.2011.11.008.

[12] L. Gebauer, J. Skewes, G. Westphael, P. Heaton, and P. Vuust, 'Intact brain processing of musical emotions in autism spectrum disorder, but more cognitive load and arousal in happy vs. sad music', *Front. Neurosci.*, vol. 8, Jul. 2014, doi: 10.3389/fnins.2014.00192.

[13] F. Y. N. Leung *et al.*, 'Emotion recognition across visual and auditory modalities in autism spectrum disorder: A systematic review and meta-analysis', *Dev. Rev.*, vol. 63, pp. 1–47, 2022.

[14] P. Heaton, 'Interval and contour processing in autism', *J. Autism Dev. Disord.*, vol. 35, no. 6, pp. 787–793, Dec. 2005, doi: 10.1007/s10803-005-0024-7.

[15] S. Stanutz, J. Wapnick, and J. A. Burack, 'Pitch discrimination and melodic memory in children with autism spectrum disorders', *Autism*, vol. 18, no. 2, pp. 137–147, Feb. 2014, doi: 10.1177/1362361312462905.

[16] M. Sharda *et al.*, 'Music improves social communication and auditory–motor connectivity in children with autism', *Transl. Psychiatry*, vol. 8, no. 1, p. 231, Dec. 2018, doi: 10.1038/s41398-018-0287-3.

[17] J. Xavier, V. Vignaud, R. Ruggiero, N. Bodeau, D. Cohen, and L. Chaby, 'A multidimensional approach to the study of emotion recognition in autism spectrum disorders', *Front. Psychol.*, vol. 6, 2015, Accessed: Jun. 09, 2022. [Online]. Available: https://www.frontiersin.org/article/10.3389/fpsyg.2015.01954

[18] G. Charbonneau *et al.*, 'Multilevel alterations in the processing of audio–visual emotion expressions in autism spectrum disorders', *Neuropsychologia*, vol. 51, no. 5, pp. 1002–1010, Apr. 2013, doi: 10.1016/j.neuropsychologia.2013.02.009.

[19] E. Kinnaird, C. Stewart, and K. Tchanturia, 'Investigating alexithymia in autism: A systematic review and meta-analysis', *Eur. Psychiatry*, vol. 55, pp. 80–89, Jan. 2019, doi: 10.1016/j.eurpsy.2018.09.004.

[20] G. Bird and R. Cook, 'Mixed emotions: the contribution of alexithymia to the emotional symptoms of autism', *Transl. Psychiatry*, vol. 3, no. 7, pp. e285–e285, Jul. 2013, doi: 10.1038/tp.2013.61.

[21] S. Baron-Cohen, S. Wheelwright, R. Skinner, J. Martin, and E. Clubley, 'The Autism-Spectrum Quotient (AQ): Evidence from Aperger syndrome/high-functioning autism, males, and females, scientists and mathematicians', *J. Autism Dev. Disord.*, vol. 31, no. 1, pp. 5–17, 2001.

[22] R. M. Bagby, J. D. A. Parker, and G. J. Taylor, 'The twenty-item Toronto Alexithymia scale—I. Item selection and cross-validation of the factor structure', *J. Psychosom. Res.*, vol. 38, no. 1, pp. 23–32, Jan. 1994, doi: 10.1016/0022-3999(94)90005-1.

[23] J. H. Ong, F. Leung, and F. Liu, 'The Reading Everyday Emotion Database (REED)'. University of Reading, 2021. doi: 10.17864/1947.000336.

[24] A. L. Anwyl-Irvine, J. Massonnié, A. Flitton, N. Kirkham, and J. K. Evershed, 'Gorilla in our midst: An online behavioral experiment builder', *Behav. Res. Methods*, vol. 52, no. 1, pp. 388–407, Feb. 2020, doi: 10.3758/s13428-019-01237-x.

[25] D. Bates, M. Maechler, B. Bolker, and S. Walker, 'Fitting linear mixed-effects models using lme4', *J. Stat. Softw.*, vol. 67, no. 1, pp. 1–48, 2015, doi: 10.18637/jss.v067.i01.

[26] J. Fox and S. Weisberg, *An R Companion to Applied Regression*, 3rd ed. Thousand Oaks, CA: Sage, 2019. [Online]. Available: https://socialsciences.mcmaster.ca/jfox/Books/Companion/

[27] R. V. Lenth, 'Least-Squares Means: The R Package lsmeans.', *J. Stat. Softw.*, vol. 69, no. 1, pp. 1–33, 2016, doi: 10.18637/jss.v069.i01.

[28] S. Baron-Cohen, A. Spitz, and P. Cross, 'Do children with autism recognise surprise? A research note', *Cogn. Emot.*, vol. 7, no. 6, pp. 507–516, Nov. 1993, doi: 10.1080/02699939308409202.

[29] F. Happé, 'An advanced test of theory of mind: Understanding by story characters' thoughts and feelings by able autistic, mentally handicapped, and normal children and adults', *J. Autism Dev. Disord.*, vol. 24, no. 2, pp. 129–154, Apr. 1994, doi: 10.1007/bf02172093.

[30] S. Baron-Cohen, A. M. Leslie, and U. Frith, 'Does the autistic child have a "theory of mind"?', *Cognition*, vol. 21, pp. 37–46, 1985, doi: 10.1097/00005373-196307000-00010.

[31] S. R. Livingstone, W. F. Thompson, M. M. Wanderley, and C. Palmer, 'Common cues to emotion in the dynamic facial expressions of speech and song', *Q. J. Exp. Psychol.*, vol. 68, no. 5, pp. 952–970, May 2015, doi: 10.1080/17470218.2014.971034.

[32] N. J. Sasson and K. Bottema-Beutel, 'Studies of autistic traits in the general population are not studies of autism', *Autism*, p. 13623613211058516, Nov. 2021, doi: 10.1177/13623613211058515.

[33] S. Olderbak, O. Wilhelm, A. Hildebrandt, and J. Quoidbach, 'Sex differences in facial emotion perception ability across the lifespan', *Cogn. Emot.*, vol. 33, no. 3, pp. 579–588, Apr. 2019, doi: 10.1080/02699931.2018.1454403.

Preliminary analysis of /r/ acoustics and features in three Māori speakers

Isabella Shields[1], *Catherine Watson*[1], *Peter Keegan*[1]

[1]University of Auckland, New Zealand

ishi836@aucklanduni.ac.nz, c.watson@auckland.ac.nz, p.keegan@auckland.ac.nz

Abstract

This paper presents preliminary analyses of Māori /r/ in the speech of three fluent speakers. Māori /r/ is not yet fully understood and can be a somewhat misunderstood sound for Māori learners. We consider /r/ in several vowel environments and word- and phrase-stress contexts. Duration of /r/ was found to vary across speakers and phrasal stress. Most /r/ tokens were taps, although variation was found in formant behaviour and presence of frication and release bursts. Formant assessment of /VrV/ sequences indicates different articulations for /r/ are possible. We emphasise the need for more speakers/participants to characterise the sound.

Index Terms: Māori language, acoustic phonetics, rhotics

1. Introduction and Background

This study presents preliminary findings of a speech study investigating the acoustic features of the /r/ phoneme in Māori. This sound is as yet not fully characterised nor understood; while Māori /r/ has been described, it has not been the subject of a comprehensive acoustic investigation. This paper outlines the methodology of an ongoing investigation of Māori /r/ and presents results based a small set of speakers. The overall aim of the study is to identify the acoustic characteristics of Māori /r/ and assess the impact that factors such as vowel environment and stress environment have on the sound.

Māori is a Polynesian language indigenous to Aotearoa (a popular Māori name for New Zealand). While Māori and New Zealand Sign Language are the only official languages of Aotearoa, English is the most commonly spoken language. As a result of the colonisation of Aotearoa in the 1800s, there has been significant language contact between Māori and English. This language contact, accompanied by active discouragement of Māori language use, resulted in a significant decline in the number of Māori speakers that eventually resulted in a break in inter-generational transmission of the language. There has been notable success in revitalising Māori, with efforts spearheaded by Māori language communities [1]. The present research aims to contribute to our understanding of the Māori language and its revitalisation, and as our research involves Māori researchers and Māori data, it must adhere to principles of Māori data sovereignty [2, 3].

1.1. Māori phonology

Māori has five monophthongs: /i e a o u/. These vowels are high-front, mid-front, open, mid-back, and high-back, respectively [1]. There are 10 consonants (/ p t k m n ŋ f h w r /), although some variation across dialects exists. The five vowels have phonemically distinctive long and short quantities, although this distinction is weakening [4]. Long vowels are usually denoted with a macron in modern written Māori. Diphthongs can be formed by combinations of these monophthongs. Māori is considered by some to be a mora-timed language, with a mora consisting of a short vowel and optional preceding consonant (μ = (C)V; C = consonant, V = short vowel) [5]. Biggs' stress rules put forward explanations for placement of both word and phrase stress [6]. For monomorphemic words, stress placement is dictated by a syllable hierarchy, with the highest ranked syllable bearing the stress. This hierarchy, as outlined by Bauer, is: $(C)V_1V_1 > (C)V_1V_2 > (C)V_1$. Biggs' rules suggest that, in a phrase that is sentence-final, phrase stress should fall on the primary word stress of the final content word. Otherwise, phrase stress should fall on the phrase's penultimate mora. For example, when the phrase [*ki te kura nei*] (meaning *to this school*) is sentence final, the phrase stress (underlined) overlaps with the stressed mora of the content word (in bold) ([*ki te* ***ku***<u>***u***</u>*ra nei*], otherwise it falls on the penultimate mora ([*ki te* ***ku****ra* <u>*nei*</u>].

1.2. Māori /r/: existing descriptions and investigations

The focus of the present study, /r/, is most often described as an alveolar flap (e.g. in [7]). However, there is some inconsistency in its description. Trilled /r/ has been reported in some older descriptions, and approximant /r/ has also been reported by Harlow [8, 1]. He notes that outside of rapid speech and repeated sequences with intervening unstressed vowels, approximant /r/ is a direct result of English pronunciation influencing that of Māori. Harlow and Biggs both refer to lateral realisations of the phoneme, but with emphasis on different geographical locations: Harlow points to South Island dialects, and Biggs to Eastern dialects [1, 8]. The spectrographic characteristics and duration of /r/ were analysed in the speech of male *kaumātua* (elders) from the MAONZE corpus [9, 10]. The findings of this study supported the canonical designation of Māori /r/ as a flap, as well as Harlow's description of the contexts in which approximant /r/ appear. The Ngā Mahi recordings, a collection of recordings of one speaker produced for the development of Māori text-to-speech system (see [11]), were also investigated in the context of lexical stress and higher formant behaviours that could not be analysed in the MAONZE corpus as a result of audio quality limitations [10, 12]. A phonetic correlate of stress reported by Bauer is 'emphatic onset', by which /r/ should increase in duration when stressed [7]. In greater proximity to a stressed syllable, /r/ duration and intensity were found to increase and decrease, respectively. For the same speaker, notable lowering of the fourth formant (F4) in a range of segmental contexts was identified as a salient feature.

2. Methodology

2.1. *Kaikōrero* (speakers)

To date, speakers have been recruited using a *kaupapa Māori* (Māori principled) approach [13]. However, the recruitment process for the present study has been slow-moving given

COVID-19 restrictions and considerations. We observe that many Māori speakers can still feel a sense of *whakamā* (shame or embarrassment) when sharing it with others and feel reluctant to do so.

There are three *kaikōrero* (speakers) analysed in the present study: Speakers 1 and 2 are a 57 year old man and a 53 year woman, while Speaker 3 is a 17 year old woman. These speakers are fluent speakers of Māori, and make up a *whānau* (family unit). They use Māori in their daily lives, both at home and at work or school. All speakers are also fluent speakers of Standard New Zealand English, with Speaker 1 additionally reporting basic competency in Malay, the canonical /r/ of which is a trill or tap, although fricatives also appear [14]. Speakers 1 and 3 reported speaking Māori with their friends. Speakers 2 and 3 both indicated they learned Māori as children (at home and on the *marae* (tribal meeting place); and at home and in immersion schooling, respectively), while Speaker 1 learned the language post-adolescence.

2.2. Materials

Five word forms were selected which place /r/ in different positions relative to word stress (WS) as predicted by Biggs' stress rules. Of these, we report on three in the present study, with syllable boundaries denoted by /./: /ˈCV.rV/ (WS1; /r/ is in the unstressed onset), /ˈrV.CV/ (WS2; /r/ is in the stressed onset and word-initial), and /ˈCV_{long}.CV.rV/ (WS3; /r/ is unstressed and placed further from the stressed syllable). Nasal consonants were avoided in the target words to avoid potential nasalisation effects on surrounding sounds. Six immediate vowel contexts for /r/ are considered: /iri, ira, iro, ara, ari, aro/. These vowel environments were selected to cover a range of environments and tongue movements, and were based on Māori point vowels. While the vowel /u/ could have been included instead of /o/, it has become increasingly fronted [15]. The target words used are shown in Table 1, with the stressed mora in bold.

Table 1: *Target words used in the elicitation study.*

WS	/iri/	/ira/	/iro/	/ara/	/ari/	/aro/
1	**pi**ri	**hi**ra	**pi**ro	**pa**ra	**pa**ri	**pa**ro
2	**ri**po	**ra**pi	**ro**pi	**ra**pu	**ri**pa	**ro**tu
3	**tā**piri	**pā**kira	**kō**piro	**hō**para	**tō**kari	**tā**karo

In order to account for any potential influences of phrase stress, two different environments for the target word were selected. Carrier sentences are used to deliver the target words, as recommended in [16, 17]. In line with [17], the target word is shielded from the end of a phrase or sentence to avoid any boundary effects. The two carrier sentences used were:

- Sentence (1): [*Ka/I* {target} *tonu*][*a* {Name}] - E.g. *Ka **tā**karo tonu a Pita* (*Pita will still play*).
- Sentence (2): [*I kite*][*a* {Name}][*i ngā* {target} *nā*] - E.g. *I kite a Tia i te **tā**karo nā* (*Tia saw that game*).

The target word functions as a verb and as a noun in Sentences (1) and (2), respectively. In Sentence (2), Biggs' rules suggest the phrase stress falls on the primary word stress of the target word. In the case of Sentence (1), Biggs' rules suggest phrase stress does not fall on the target word, but rather on the penultimate mora of the phrase. Each sentence was randomly assigned one of six two-syllable Māori first names. In order to account for any temporal variations in speakers' production, the sentence list was repeated five times and its order randomised during each repetition. To add variation, a dummy sentence was added to the sentence list every 10 sentences.

It was not possible to formulate sentences which maintained the initial /i/ of /iri, ira, iro/ in WS2 words in Sentence (2) while also maintaining syntactically meaningful sentences. Instead, the plural article *ngā* (*the*; /ŋaː/) was replaced with the singular *te* (*the*, /te/), producing the sequences /eri, era, ero/. While not directly equivalent, these sequences were the closest approximation possible.

2.3. Recording process and speech task

Recordings for this study were approved by the University of Auckland Human Participants Ethics Committee (UAHPEC23198). Speech recordings were made using a Rode Lavalier lapel microphone and Roland OCTA-CAPTURE for preamplification and digitisation. Audio was captured at a sample rate of 44.1kHz and 16-bit bit depth. A maximum of five repetitions of the sentence lists were recorded, depending on the speed of the speaker. Digital capture and management of the display of recording materials was achieved using Python (version 3.7.2) and the *sounddevice* package (version 0.4.4) [18]. Recordings took place in a WhisperRoom Sound Isolation Enclosure (`https://whisperroom.com`), with speakers undertaking the task seated at a desk in front of a computer monitor. 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. Speakers 1 and 3 completed the entire speech task and Speaker 2 completed four of the five repetition rounds within the hour.

2.4. Data preparation

Data processing, analysis, and visualisation was undertaken using R (version 4.2.0) [19]. Recordings were converted into speech databases using the package *EmuR* (version 2.3.0) [20]. Initial phonetic segmentation of recordings was achieved using WebMAUS General (language: language independent (SAMPA), otherwise default settings) [21, 22, 23]. Boundaries were then hand-corrected where necessary in *EmuR*. Start and end boundaries for /r/ 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, and then hand-adjusted by the first author [24]. When a formant was weak enough to not be visible in a spectrogram, the formant value was excluded. This was particularly common for the fourth formant (F4), especially in the case of Speakers 2 and 3. All formant energies were also frequently not visible throughout the /r/.

2.5. Feature labelling

Labels for spectrographic/waveform features were added based on visual inspection of each /r/ token. These described the presence of a release burst ('B'), the presence of at least one formant between F1 and F4 throughout the /r/ ('R'), and the presence of frication in at least part of the /r/ ('F'). All cases of frication present were identified as a secondary articulation with the same primary manner of articulation (flap). In the rare cases where a different primary manner of articulation was identified (such as

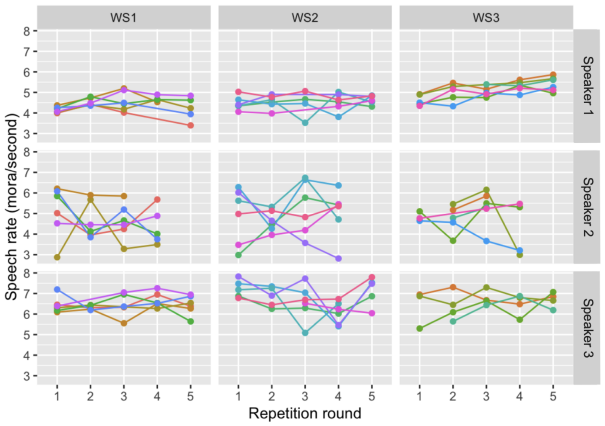


Figure 1: *Changes in speech rate (mora/s) for Speakers 1-3 in Sentence (1). Each colour corresponds to a different sentence.*

an approximant or trill), the tokens were labelled accordingly.

3. Results

In total 475 /r/ tokens were analysed. A total of 22 tokens were excluded from analysis, with exclusions made when the speaker made an error, the production was breathy, or where noise from movement interfered with the recording quality.

3.1. Speech rate

We considered speech rate on a speaker-by-speaker basis to determine if there was variation throughout the speech task speech Speech rate was calculated in terms of morae per second, with the number of morae in each sentence determined manually and divided by the duration of each utterance as extracted from the corrected EmuR database. The speakers did not show a noticeable tendency to increase/decrease their speech rate during the speech task, save for in Speaker 1's productions of Sentence (1)/WS3 sentences, which showed a slowly increasing speech rate as the task progressed. Speech rate for each repetition of each Sentence (1) sentence is shown in Figure 1. These results are comparable to Sentence (2) results. Visual inspection of Figure 1 shows Speaker 2, with a range of 3.9 morae, demonstrated more variation in speech rate across the different sentences and repetitions than Speakers 1 and 3 who had a speech rate range of 1.7 morae and 2.7 morae, respectively. This may indicate unease with some aspect of the experiment. Speaker 2 did report that she felt Sentence (2) sentences were odd to speak aloud due to the final word (*nā*, a locative particle). Speaker 3 tended to produce the sentences at a higher speech rate. This may be a result of comfort with the speech task or age, with younger speakers reported to have higher speech rate [25].

3.2. /r/ type and spectrographic features

All instances of /r/ observed were taps/flaps, save for three examples of an approximant and one trill. We conclude that the approximant and trill observations were mostly likely speaker error. Furthermore, no lateral productions of /r/ were observed. A majority of tokens (91.8%) had at least one formant (F1-F4) visible throughout. Frication was identified in 45.3% of observations, and release bursts in 12.6% of observations. On a speaker-by-speaker basis, the occurrence of these features was largely similar. Speaker 2 produced a release burst in 25.4% of their /r/ productions, compared to 6.5% and 8.8% for Speakers 1 and 3, respectively. Instances of /r/ with frication were observed in around half of Speaker 1 and 2 tokens (50.0%, and 49.3%, respectively), but in 37.4% of Speaker 3 tokens. We also note that tokens with frication appeared more frequently in the vicinity of a front vowel than not: frication appeared in 58.3% of tokens preceded/followed by a front vowel, compared to 19.5% of tokens in solely open/back environments.

3.3. Duration of /r/

Table 2 summarises the mean and standard deviation duration of /r/ tokens for word stress and sentence environment. Overall, mean /r/ duration was observed to be greater in the Sentence (2) environment than in Sentence (1), albeit to differing degrees. In Sentence (1) and (2), both Speaker 1 and Speaker 2 produced /r/ with greater mean duration in the WS2 context than in WS1 or WS3. These observations appear to support the observations made by Bauer regarding increased /r/ duration when in the onset of the stressed syllable [7]. Speaker 3 did not follow this trend, with the WS3 context (where /r/ is not in the onset of the stressed syllable) having the greatest mean /r/ duration in both Sentence (1) and (2). We also note there was a tendency for /r/ duration to increase when in the vicinity of a front vowel (such as in the sequence /iri/ or /ira/), but not in the sequence /iro/. As indicated above, /r/ tokens with frication occurred more frequently in these sequences than in others.

Table 2: *Mean duration and standard deviation of /r/ (ms) by stress environments (n>21 samples in each category).*

Mean /r/ duration and S.D. (ms) in Sentence (1)			
Speaker	WS1	WS2	WS3
1	29.0 (6.9)	35.3 (8.8)	29.3 (6.5)
2	24.3 (7.2)	35.2 (11.8)	28.8 (7.8)
3	26.5 (10.4)	26.7 (8.0)	30.2 (8.9)
Mean /r/ duration and S.D. (ms) in Sentence (2)			
1	30.0 (4.1)	36.7 (7.4)	33.3 (5.2)
2	29.8 (7.3)	41.8 (16.5)	37.7 (10.0)
3	29.8 (11.8)	29.9 (11.8)	37.1 (12.3)

3.4. Formants

The formant trajectories of the vowels preceding and following /r/ in the target words were considered in each sentence and word-stress environment. For the three speakers analysed, there did not appear to be any notable visual discrepancies between the formant trajectories of the two phrase stress environments, save for greater variation in estimated formant values in the Sentence (2) environment. In Figure 2 we include formant trajectories of /ari/ ((a), (c), and (e)) and /iro/ ((b), (d), and (f)) sequences in the Sentence (1)/WS1 environment, however the present discussion is not limited to these sequences.

The movements of the first and second formants (F1 and F2) provided some insight into the movements of the tongue before and after /r/ was articulated, and indicated there are different articulatory approaches to the various /VrV/ sequences. Preparatory movement associated with /r/ appears to occur to differing degrees depending on the /VrV/ sequence. For example, in the sequence /ari/ (pictured in Figure 2 (a), (c) and (e)), Speaker 2 and 3 show a steady raising of F2 accompanying the forward movement of the tongue towards the /r/ target, followed

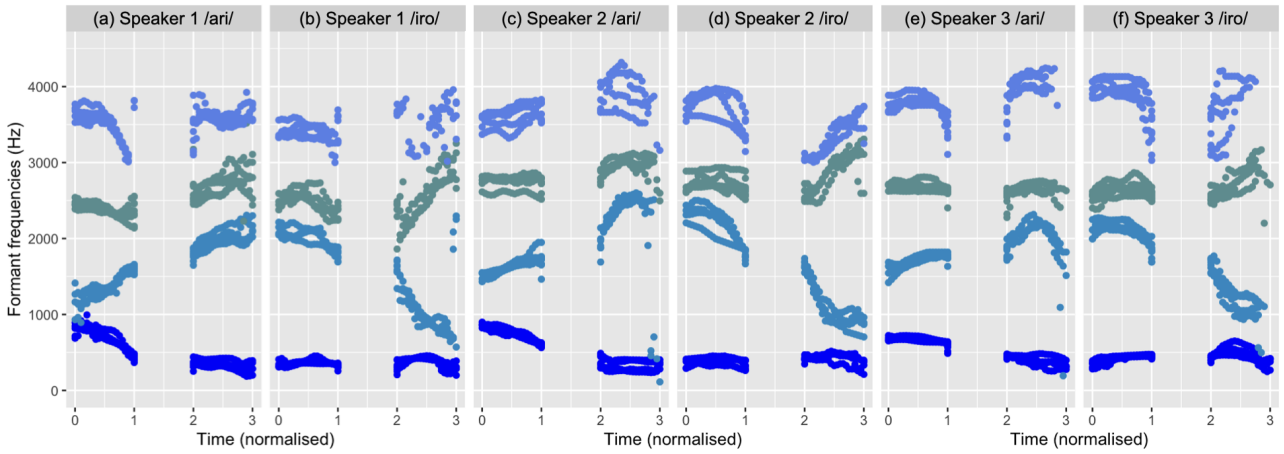


Figure 2: *Formant trajectories (F1-F4) of vowels preceding and following /r/ in /ari/ and /iro/ sequences in WS1/Sentence (1). Time 0-1 shows the formant trajectories of the preceding vowel, Time 1-2 corresponds to the /r/, and Time 2-3 to the following vowel.*

by more abrupt movement towards the /i/ target. Conversely, F2 for Speaker 1 appears to change at much the same rate both before and after /r/. Relatively little movement in F1 in observations of /iro/ demonstrated there was not much change in tongue height in the transition from /i/ to /r/, and /r/ to /o/. Conversely, there was notable lowering of F1 for Speaker 1 and 2 in the /a/ of /ari/, indicating gradual raising of the tongue to reach the alveolar ridge to produce /r/, before remaining at much the same height in preparation for articulating /i/.

Inference regarding articulation based on the third and fourth formants (F3 and F4) was more complicated. Similar to F1 and F2, preparatory changes leading into /r/ began at various points in the preceding vowel. Lowering of F3 is often associated with rhotic sounds [26]. For all speakers, we note /r/ accompanied by lowered F3 occurred in the presence of /o/. Māori /o/ is lip-rounded, contributing to the lowering of F3 (and the other formants). Outside of these contexts, F3 lowering was observed in various segmental environments for Speaker 1, less frequently for Speaker 2, and sporadically for Speaker 3. Notable movement of F4 preceding /r/ (lowering prior to /r/ and/or raising following /r/) occurred across all vowel environments for Speaker 1 and 2, albeit to differing degrees. For Speaker 3, visually salient lowering of F4 appeared to only occur in /iro/ (see Figure 2 (f)), /ero/, and /aro/ vowel environments. The source of this F4 lowering is not immediately evident, and may indicate a particular articulatory approach to /r/ in certain environments. This feature will be investigated further in future.

4. Discussion and Conclusions

The present study aimed to identify some of the acoustic characteristics of Māori /r/ and to investigate the potential impacts of segmental and stress contexts. To our knowledge, this study is the first of its kind formulating and analysing a speech task focusing on Māori /r/. Given the study is ongoing and we are presently reporting on a small number of speakers, we have focused on description rather than statistical analyses.

Inter- and intra-speaker variation is a common trait of rhotic sounds in several languages (such as Portuguese [27], Dutch [28], German [29], and Spanish [30]). In the present study, we find relatively little variation in /r/ type, but variation in other features like frication, presence of release bursts, duration, and formant behaviours. While we did not observe a consistent indication that word stress was influencing /r/ duration, it was generally shown to increase in Sentence (2) contexts. In Sentence (2), Bigg's stress rules suggest the phrase stress falls on the same mora as the word stress of the target word, placing more prominence on the mora compared to Sentence (1). It is possible that phrase stress is indeed influencing the duration of /r/, or that some combined interaction of phrase stress and word stress results in increased /r/ duration.

The majority of /r/ tokens (91.8%) had formant energy present throughout, indicating that the closure in the oral cavity produced during flap articulation is usually not sufficient to fully block airflow. This is not uncommon in other languages with flap sounds (e.g. American English dialect's flap allophone [31]). We report a rate of frication (45.3% of tokens) that is surprisingly high given it is not often mentioned in existing descriptions of Māori /r/. Given the majority of these observations occurred near a front vowel, we speculate there may be an interaction between /r/ and this environment by which the proximity of the articulatory targets of /i/ and /r/ results in slightly reduced contact of the tongue and alveolar ridge in the flap movement, allowing for a sustained flow of air with introduced turbulence.

In addition to further speaker recordings, an articulation-focused investigation would be useful to determine the exact source of frication alongside further acoustic investigation, and would also tease out complex articulatory information not easily inferred from higher formant behaviours (i.e. F3 and F4).

4.1. Conclusions

Detailed investigation of the acoustics of Māori /r/ is necessary if the sound is to be well understood. The preliminary results presented here represent the first steps in achieving this goal. Our findings contribute to the wider understanding of the characteristics of /r/; in particular, we identify variation in its duration in different stress and segmental environments, as well as indication of different articulatory approaches to producing the sound based on acoustic features and formant behaviours. We hope that with further investigation the source and breadth of these variations will become evident. The results of this preliminary investigation indicate there is a need to further home in on these various contexts with a wider speaker/participant group to enable a more precise description of /r/ in modern Māori.

5. References

[1] R. Harlow, *Māori: A Linguistic Introduction*, Cambridge, U.K.: Cambridge University Press. 2007.

[2] T. Kukutai and J. Taylor, Eds., *Indigenous Data Sovereignty: Toward an agenda*. Canberra, Australia: ANU Press, 2016.

[3] M. Hudson, T. Anderson, T. K. Dewes, P. Temara, H. Whaanga, and T. Roa, "'He Matapihi ki te Mana Raraunga' - Conceptualising big data through a Māori lens," in *He Whare Hangarau Māori: Language, culture & technology*, H. Whaanga, T. T. Keegan, and M. Apperley, Eds. Kirikiriroa/Hamilton: Te Whare Wānanga o Waikato/University of Waikato, 2017, pp. 64–73.

[4] J. King, R. Harlow, C. Watson, P. Keegan, and M. Maclagan, "Changing pronunciation of the Māori language: implications for revitalization," *Indig. Lang. Revital. Encourag. Guid. lessons Learn.*, pp. 85–96, 2009.

[5] W. Bauer, "Hae.re vs. ha.e.re: a note," *Te Reo*, vol. 24, pp. 31-36, 1981.

[6] B. Biggs, *Let's learn Maori: a guide to the study of the Maori language*, 3rd ed., Auckland, N.Z.: Auckland University Press, 1998.

[7] W. Bauer, *Maori*, London/New York, N.Y., U.K./U.S.A.: Routledge, 1993.

[8] B. Biggs, "The Structure of New Zealand Maaori," *Anthropological Linguistics*, vol. 3, no. 3, pp. 1-54, Mar., 1961. [Online]. Available: https://www.jstor.org/stable/30022302

[9] J. King, Maclagan, M., Harlow, R., Keegan, P., Watson, C. 2011. The MAONZE Project: Changing uses of an indigenous language database. *Corpus Ling. and Ling. Theory*, vol. 7, no.1, pp. 37-57, Apr., 2011.

[10] I. Shields, C. Watson, P. Keegan, and M. Maclagan, "A preliminary investigation of the acoustics of Māori /r/," New Zealand Linguistics Society Conference, Hamilton, New Zealand, Dec., 2021. [Presentation]

[11] I. Shields, C. Watson, P. Keegan, R. Berriman and J.James "Creating a Synthetic Te Reo Māori Voice," International Conference on Language Technologies for All, Paris, Dec 2019. [Online]. Available: https://lt4all.org/media/papers/P1/136.pdf.

[12] I. Shields, C. Watson, P. Keegan, "Acoustics of te reo Māori /r/: First impressions and corpus building for fourth formant and stress analysis," 'r-atics 7 Conference, Lausanne, Switzerland, Nov., 2021. [Presentation].

[13] L. T. Smith, *Decolonizing Methodologies: Research and Indigenous Peoples*, 3rd ed., London, U.K.: Zed Books. [Online]. Available: https://bloomsbury.com/9781786998132.

[14] D. Deterding, I. A. Gardiner, and N. Noorashid, *The Phonetics of Malay*, Cambridge, U.K.: Cambridge University Press, 2022.

[15] 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*, vol. 21, no. 2, pp. 175-192, Jul., 2009, doi: doi:10.1017/S095439450999007X.

[16] J. Harrington, *Phonetic Analysis of Speech Corpora*, Chichester, U.K.: Wiley-Blackwell, 2010.

[17] T. Roettger and M. Gordon, "Methodological issues in the study of word stress correlates," *Linguist. Vanguard*, vol. 3, no. 1, 2017, doi: 0.1515/lingvan-2017-0006.

[18] M. Geier. *sounddevice*. [Online]. Available: https://github.com/spatialaudio/python-sounddevice.

[19] R Core Team (2022). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL: https://www.R-project.org/.

[20] R. Winkelmann, K. Jaensch, S. Cassidy and J. Harrington. *emuR: Main Package of the EMU Speech Database Management System*, R package version 2.3.0.

[21] F. Schiel. "Automatic Phonetic Transcription of Non-Prompted Speech," in Proc. of the ICPhS, San Francisco, C.A., U.S.A., 2015, pp. 607-610.

[22] F. Schiel. "A Statistical Model for Predicting Pronunciation," in Proc. of the ICPhS 2015, Glasgow, U.K., 2015, paper 195.

[23] T. Kisler, U. D. Reichel and F. Schiel, "Multilingual processing of speech via web services," *Computer Speech & Language*, vol. 45, no. 1, pp. 326–347, Sept., 2017.

[24] L. Bombien, R. Winkelmann and M. Scheffers. *wrassp: an R wrapper to the ASSP Library R package*, R package version 1.0.1.

[25] J. Bóna,"Temporal characteristics of speech: The effect of age and speech style," *J. Acoust. Soc. Am.*, vol. 136, no. 2, pp. EL116-EL121, Aug., 2014, doi: 10.1121/1.4885482.

[26] P. Ladefoged and I. Maddieson, *The sounds of the world's languages*, 1st ed. Oxford/Cambridge, M.A., U.K./U.S.A.: Blackwell, 1996.

[27] I. E. Rennicke, "Variation and Change in the Rhotics of Brazilian Portuguese," Ph.D. dissertation, Faculty of Arts, Federal University of Minas Gerais, Belo Horizonte, Brazil, 2015.

[28] K. Sebregts, "The Sociophonetics and Phonology of Dutch r," Ph.D. dissertation, University of Utrecht, Utrecht, Netherlands, 2015.

[29] R. Wiese, "The Unity and Variation of (german) /r/," *Zeitschrift für Dialektol. und Linguist.*, vol. 70, no. 1, pp. 25–43, Jul. 2003, [Online]. Available: http://www.jstor.org.ezproxy.auckland.ac.nz/stable/40504887.

[30] E. W. Willis, "An acoustic study of the 'pre-aspirated trill' in narrative Cibaeño Dominican Spanish," *J. Int. Phon. Assoc.*, vol. 37, no. 1, pp. 33–49, 2007, doi: 10.1017/S0025100306002799.

[31] N. Warner, A. Fountain, and B. V Tucker, "Cues to perception of reduced flaps," *J. Acoust. Soc. Am.*, vol. 125, no. 5, pp. 3317–3327, 2009, doi: 10.1121/1.3097773.

Towards the automatic identification of /l/-vocalisation in English speakers in Australia

Simon Gonzalez[#1], *Gerard Docherty*[*2]

[#]Australian National University, Australia
[*]Griffith University, Australia
[1]u1037706@anu.edu.au, [2]gerry.docherty@griffith.edu.au

Abstract

The aim of this paper is to describe the initial development of a computational framework designed to automatically recognize and classify vowel-/l/ rhyme realisations produced by Australian English speakers as either consonantal or vocalised. We implemented a Random Forest model as the main classificatory technique. This allowed us to explore in a hierarchical way the contribution to the classification of a wide a range of potential predictors. The test classification accuracy of the Random Forest model was 82.1% overall, with its sensitivity estimated to be 73.7% (consonantal realisations) and the specificity to be 89.1% (vocalised realisations).

Index Terms: /l/-vocalisation, machine learning, Australian English, sociophonetics

1. Introduction

Advances in forced alignment and segmentation have transformed the toolkit available to researchers interested in tracking the variable performance of speakers in large-scale corpora of natural speech. The spectral and time-domain properties of vowels are well-understood and key parameters such as formant frequencies can be extracted automatically from corpora using off-the-shelf tools such as FAVE [1] or MAUS [2]. In recent years progress has been made in automatically extracting some of the variable realisational properties of consonants (e.g. Voice Onset Time (VOT) [3] and the spectral properties of fricative realisations [4]). Some phonetic properties of speech, however, present greater challenges for automatic identification and classification, especially in spontaneous speech styles.

One such property is the variable realisation of post-vocalic /l/ across different varieties of English. In syllable-coda position (most notably in pre-pausal position or in the context of a following syllable that begins with a consonant) /l/ can be realised as a canonical lateral approximant with a central occlusion of the oral cavity and uni- or bi-lateral airflow. With concurrent voicing, this gives rise to a spectral signature not unlike that of a vowel segment, but typically with discontinuities associated with the formation of the central occlusion. However, if the occlusion for a coda /l/ is not fully formed, the /l/ is said to have a vocalised realisation and its auditory and acoustic properties are rendered much more vowel-like with the result that it can be hard to distinguish a vocalised /l/ from back rounded vowels (which have a very similar articulatory configuration).

This phenomenon has been reported in a number of languages across the world (e.g. Swiss German [5], Dutch [6], and Portuguese [7]). In English, /l/-vocalisation is a widely observed phenomenon across many varieties [8] [9] [10], including those spoken by speakers in Australia [11] [12] where its occurrence has been found to be influenced by speaker age and provenance, as well as by a range of context-dependent factors. For example, as well as finding differences across different cities and age-groups, Horvath and Horvath [11] found that high vowels are more likely to trigger /l/-vocalisation than low vowels, and that back vowels are more likely to be associated with vocalisation than front and central vowels. The same study reported a greater likelihood of vocalisation following phonologically long monophthongs than following diphthongs and short monophthongs, and differences as a function of the place of articulation of an adjacent consonantal context. It was also found that tokens followed by a consonant were more likely to be vocalised than those occurring in pre-pausal and pre-vocalic contexts. In general, however, it is fair to say that this is a rather under-documented feature of the phonetic variability characterising English speakers in Australia.

Experimental studies of the occurrence of /l/-vocalisation in English have largely made use of auditory [13][14] and/or articulatory methods [15][16][17] [18][19]. These studies bring into question the nature of the phenomenon in itself, pointing to vocalisation being less obviously a categorical process as opposed to being located at one end of an articulatory light-dark articulatory/auditory continuum, albeit a continuum that seems to embed an auditory discontinuity given researchers' confidence in detecting at least some V-/l/ tokens as clearly vocalised [14], with the latter sort of auditory judgement constituting the basis for most work to date on the socio-phonetic/-linguistic aspects of this phenomenon.

This represents quite a challenge for sociophonetic research which generally seeks to quantify the occurrence of discrete variants within large speech corpora across speakers, styles, locations, or to track gradient acoustic measures that are known to be valid metrics (e.g. formant values, VOT, or spectral Center of Gravity (CoG) for fricatives) of a particular gradient realisational variant. In order to achieve this for /l/-vocalisation, what would be ideally be required is an acoustic measure mapping to the articulatory continuum allowing investigators to test for the occurrence of any socially-correlated variation within the gradient realisation of coda /l/, but in reality, for now, investigators remain largely reliant on a by-hand auditory analysis that seems ill-suited to the phenomenon which it is attempting to capture (leading [14] to note that "the precise phonetic difference between velar (L) and vocalized (L) is one of the more subtle variable distinctions in sociophonetic research and presents one of the biggest methodological challenges").

In this study, we investigate the possibility that a machine learning method might provide a means of addressing some of

the methodological challenges associated with vocalised /l/. The question addressed is whether there are patterns in the signal that can differentiate even conservatively-defined vocalised/non-vocalised /l/s that can be learned and applied automatically with a reasonable degree of effectiveness? If this proves to be the case, it would potentially offer a means of enabling researchers to extend the prevalent binary classification of /l/-vocalisation to larger corpora. While our study is not designed to resolve the question of whether /l/-vocalisation is best characterised as a binary variant or as one manifestation of a more complex realisational continuum, it does nevertheless provide a means of gauging if there is mileage in pursuing an automatic approach to the binary auditory judgement that predominates in this line of research.

2. Model Development

The approach adopted in this study drew on pre-existing reports of the application of machine learning methods in automatically classifying a range of different types of realisational variant, such as clear or dark /l/ in speakers of English from the USA [20], and variants of post-vocalic /r/ and medial /t/ in NZE [21]. In the approach described below, a statistical model is first trained to map between a set of acoustic parameters and the vocalised/non-vocalised labels that are to be classified. Subsequently the effectiveness of the model to classify new material is tested in order to gauge the predictive power and effectiveness of the statistical model [22]. To our knowledge, this is the first application of this type of machine learning methodology to the classification of /l/-vocalisation. This particular realisational variation is a good testbed for this approach to phonetic classification, given the potential that the specific classification method adopted has to deal with a wide range of candidate predictive parameters.

2.1. Creation of a test dataset

This study is part of a larger project investigating sociophonetic variability in the performance of English speakers from Perth in Western Australia. The dataset around which our analytic approach was developed and tested therefore comprised tokens of post-vocalic /l/ produced by young speakers of West Australian English (from the Perth metropolitan area). Recordings were obtained from two different speech styles: word-lists (12 native speakers of West Australian English – 6 females and 6 males, aged 18-22, producing a 165 item wordlist designed to elicit citation form/carefully produced tokens of key variables) and conversational speech (2 * 4 same-sex speakers recorded while participating in unscripted dyad conversations of around 30 minutes in duration).

Audio files were first segmented in Elan [23] and subsequently force-aligned within LaBB-CAT [24], using CELEX [25] and HTK [26] with manual correction of alignments. Using Praat [27], we extracted formant estimates (F1, F2, and F3), intensity, F0, and Mel Frequency Cepstral Coefficient (MFCC) tracks for each speech recording.

In order to create a dataset that could be used for benchmarking, training and testing the classification methodology, we selected from the above material 334 tokens of /l/ preceded by monophthongs and followed by either a pause or a consonant. For the purposes of this initial development phase of our work, we excluded tokens where /l/ was followed by an approximant, vowel or other liquids, in order to avoid dynamic formant trajectories beyond the offset of /l/. Table 1 summarises the characteristics of this reference dataset.

Table 1: *Count of all tokens in the reference dataset including the benchmark type of realisation as determined by the trained listeners' auditory classification*

Speech Style	Gender	Realisation	%	Count
Word List	Female	Vocalised	**13.4**	**45**
		Consonantal	**16.4**	**55**
	Male	Vocalised	**13.8**	**46**
		Consonantal	**16.2**	**54**
Conversation	Female	Vocalised	**17.4**	**58**
		Consonantal	**9.6**	**32**
	Male	Vocalised	**10.2**	**34**
		Consonantal	**3**	**10**
Total			**100**	**334**

A perceptual task was carried out to create the benchmark classifications necessary for evaluating the automatic coding framework, i.e., whether the automatic prediction from the machine learning model matches with the corresponding type of /l/-realisation. For this task three phonetically and sociolinguistically trained listeners coded each of the tokens containing the vowel+/l/ rhymes. Each rater listened to all of the tokens played in random order, and classified them into one of the following categories: Consonantal /l/, Intermediate, or Vocalised /l/ [28]. Where two of the three raters agreed on a particular classification, the majority view was adopted. If raters had no agreement on their auditory classification, all three raters consulted and reached a consensus decision after further listening to the token concerned. For the purposes of this initial test of the automatic classification methodology, we grouped the Consonantal and Intermediate classifications into one single group, following [28], thereby provide a binary-classified benchmark reference dataset, with the Vocalised set comprising those tokens that were most readily classified auditorily as containing vocalised /l/ variants.

Figure 1 presents /l/-realisation rates as coded in the auditory test pooling across the two speech styles investigated. It shows that all speakers vocalise /l/, but with substantial inter-speaker variation, ranging from speakers who mainly vocalize to speakers who mainly produce canonical realisations.

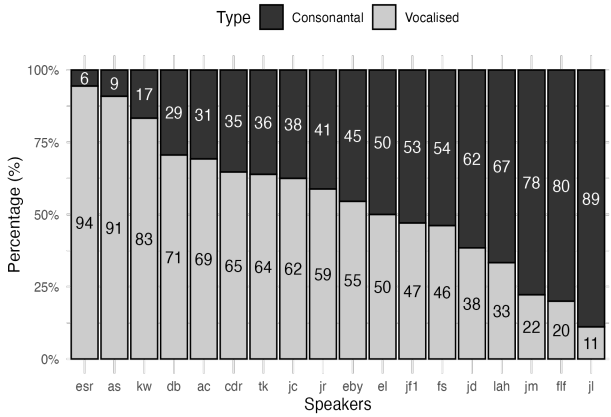


Figure 1: *% /l/-realisation type rates per speaker as per auditory test pooled across speech styles*

As has previously been noted in the literature (see above), the patterns of /l/-realisation are influenced by the identity of the vowel component of the rhyme (as shown in Figure 2). Back and central vowels (GOOSE, FOOT, THOUGHT, NURSE, SCHWA) are associated with relatively high levels of vocalised realisations across the two speech styles (bottom lighter colors), whereas front and low vowel contexts yield relatively fewer tokens of vocalisation. There are also differences across speech styles for some vowel contexts, with a relatively higher level of vocalisation found in word-list style for low and front vowels but a more even distribution across styles for the back and central vowel contexts.

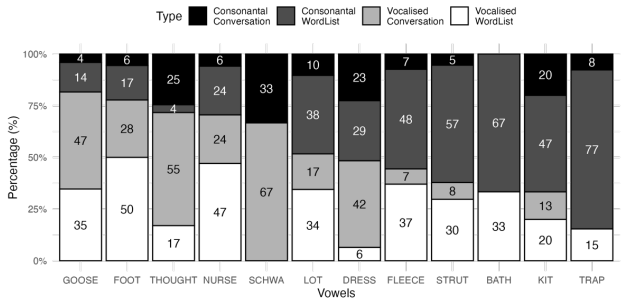


Figure 2: *% /l/-realisations by Vowel and Speech Style. The top two darker colours are consonantal realisations and the bottom two lighter colours are vocalised realisations.*

2.2. Parameterisation of the data

Following [29] and [30], for all of the selected vowel+/l/ sequences, we captured acoustic data from the onset of the vowel to the offset of the /l/. For each vowel+/l/ rhyme token the acoustic parameters referred to above (formants, intensity, F0 and MFCCs) were extracted at 11 equidistant points through an R [31] script developed for this purpose. After this, the dynamic trajectory of each acoustic measure was parameterised using a Discrete Cosine Transformation (DCT) of the 11-point trajectories. This then gave for each token a set of single values (the DCT coefficients) characterizing the trajectory for each acoustic parameter. The DCT output consists of C0-C3. C0 captures the mean amplitude, C1 the linear slope (whether it is flat or not), C2 the curvature, and C3 the amplitude at higher frequencies. This allows the learning algorithm to test which aspect of the signal may have the strongest predictive power in the classification task being undertaken. For all trajectories, the DCT transformation also smoothed out point extraction errors from the signal. A set of non-acoustic parameters accessible from the forced alignment were also built into the model including factors such as quality of the preceding vowel (on a front-back, close-open dimension), and information regarding the place and manner of articulation and voicing of adjacent consonants. This multi-faceted parameterisation of the vowel+/l/ dataset generated many potential predictors, acoustic and non-acoustic. The learning task was to determine which if any provided an effective basis for classifying tokens as vocalised or not.

2.3. Predictor Optimisation

The modelling was carried out in training, testing and assessment phases. In the training phase, given our initial broad set of variables extracted from the data (including speaker characteristics, vowel information, phonological context, and acoustic features, as described above), we applied the Boruta feature selection algorithm [32] as a Variable Importance Measure designed to identify the features of a dataset most relevant to the classification task. We ran the Boruta package in R [31], implementing a feature selection algorithm that identifies the relevant variables in a dataset [33] by iteratively removing the features which are less relevant for classifying the data. The power of this approach lies in its application to datasets in which a large number of variables need to be modeled in the absence of prior knowledge regarding which factors are likely to be the most relevant. Another crucial reason is that it has been demonstrated that there are machine learning algorithms that show a decrease in accuracy when the number of variables is higher than optimal [34]. The Boruta analysis pointed to 11 variables which were most important in distinguishing between the categories to be classified, as shown in Table 2.

Table 2: *Important parameters from the Boruta algorithm (Gini Importance). The ** represent those parameters that were not significant in the final Random Forest Model.*

Variable	Description	Boruta Importance
MFCC 5 c1	DCT Slope (coeff. 5)	**21.2**
F1 (50%)	F1 value at 50%	**20.6**
Intensity c0	DCT Mean amplitude	**20.5**
MFCC 5 (80%)	MFCC coeff. 5 at 80%	**19.6**
Intensity (20%)	Intensity at 20%	**19.2**
MFCC 8 (80%)	MFCC coeff. 8 at 80%	**16.7**
Advancement	Vowel advancement	**5.7**
MFCC 2 (20%)	MFCC coeff. 2 at 20%	**4.2**
MFCC 7 c2	DCT Curvature (coeff. 7)	**
Duration (ms)	Duration of trajectory	**
F1 c0	DCT Mean amplitude	**

2.4. Classification model training and testing

The 11 variables referred to above were chosen to be input parameters in a Random Forest (RF) model [35] implemented by using the randomForest package in R. This runs a combination of tree predictors, with each tree randomly selecting multiple predictive variables and evaluating their contribution to the classificatory task. In the end, all variables are ranked and the ones with lower error rates are classified as more important for classifying the categories specified. For this, the data was split into a training dataset and a test dataset used for validation and prediction. For the training dataset, we randomly selected 75% of the full dataset. For the test dataset, we selected the remaining 25%. In the first stage, the RF model is trained using the training dataset. Then in a second stage, the algorithm tries to predict the /l/-realization type on the test dataset, which it has not seen before. With this, we avoid overfitting on the predictability power of the model if we only test accuracy on data that the predictor has already seen.

The RF model was run using a classification method to predict the benchmark 'correct' classification of the /l/-realisation (i.e. the outcome that the RFs aims to predict)

provided by the auditory assessment. The results from this model showed that out of the eleven variables initially identified in the parameter optimisation stage of our analysis, eight proved to be important variables when running the classification, as shown in Table 2.

2.5. Estimating model effectiveness

The RF modelling was run with 500 trees, with the number of variables tried at each split (mtry) evaluated at 2. The overall test classification accuracy of the RF model was 82.1%. The correct classification of consonantal types (sensitivity) had 73.7% accuracy; i.e. over 7/10 Consonantal realisations were correctly classified as true positives by the model (with 3/10 as false negatives), while the correct classification of vocalised types (specificity) had 89.1% accuracy; i.e. almost 9/10 of the Vocalised realizations were correctly classified (with 1/10 as a false negative). The effectiveness of the model classification was somewhat variable across conditions; the most effective overall classification (87%) was found following a front vowel nucleus – probably an environment where the spectral differentiation of the two variants is relatively high.

3. Discussion

The integration of machine learning and linguistic analysis in applications such as automatic speech recognition has demonstrated that robust phonetic classification models can be developed with a great degree of accuracy. In this work, we have developed a promising model for applying these techniques to the automatic classification of /l/-realisational variants in speakers of one variety of English, an approach that approximates the 'conservative' categorical judgements of /l/-vocalisation that have prevailed in the sociophonetic literature to date. While caveats of course apply to our findings (not least in relation to the relatively small-scale benchmark dataset that we have deployed), our results suggest that there would be value in further refining this approach. A first step would be to test the modelling framework with a significantly larger benchmark dataset. It would also be important to further explore ways of optimizing the statistical model and the criterial acoustic parameters, with a view to reducing false positive classification of vocalized tokens, and we need to systematically assess reasons why classification effectiveness varies across different phonological contexts. In pursuing the latter question, one challenge is perhaps the somewhat opaque mapping between MFCC representations and the acoustic/articulatory space that we are most familiar with in accounting for contextual and coarticulatory variation in speech performance [36].

A further limitation of our approach is that it is predicated on a binary categorization of /l/-realisations and therefore does not lead us any closer to automatically capturing from large corpora socially-correlated variability in the realisation of the broader /l/-colouring continuum (of which vocalisation is most likely just one aspect). Having said that, as pointed out in [14], a binary classification of this phenomenon does seem to map to a certain extent to listener's auditory judgements of the vocalised /l/, so there is some merit in attempting to achieve automatic coding of this variable that approximates that which listeners can provide. This exemplifies a more general point of debate within sociophonetic research; namely a tension between insights gained from acoustic and/or articulatory investigations with methods that can identify deep layers of fine-grained cross-speaker variation, and other insights gained from auditory analysis based on less granular categories. This is a tension that has been captured by Labov's [37] concern about an "endless pursuit of detail", but ultimately it is a tension that underscores the need to investigate sociophonetic variability from the point of view of both speakers and listeners.

4. Conclusions

Our findings demonstrate that a Random Forest approach to modelling, deploying a range of acoustic and other parameters together with the application of a Variable Importance Measure, is a promising basis for undertaking automatic acoustic coding of vocalised /l/-realisations by speakers of Australian English. Future development of this approach will explore the relative effectiveness of the model in different phonological contexts, as well as with a much larger dataset than that which was used here in this exploratory proof-of-concept study. There is no doubt that /l/-vocalisation warrants much more investigation across many varieties/languages (including the many other varieties of English in Australia and NZ). The work reported above suggests that, with effective further refinement, automatic classification methods have potential to form part of the toolkit for this task.

5. Acknowledgements

This research was undertaken with support from the Australian Research Council (DP130104275). We also wish to thank three anonymous reviewers for their helpful comments on the shape and content of the submission.

6. References

[1] Rosenfelder, I., Fruehwald, J., Evanini, K., Seyfarth, S., Gorman, K., Prichard, H. and Yuan, J., "FAVE (Forced Alignment and Vowel Extraction) Program Suite v1.2.2", 2014.

[2] Strunk, J., Schiel, F. and Seifart, F., "Untrained forced alignment of transcriptions and audio for language documentation corpora using WebMAUS", in N. Calzolari, K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, A. Moreno, J. Odijk and S. Piperidis [Eds], Proceedings of the Ninth International Conference on Language Resources and Evaluation, LREC'14, 3940-3947, Reykjavik, Iceland, 2014.

[3] Arias-Vergara, T., Arguello-Velez, P., Vásquez-Correa, J.C., Nöth, E., Schuster, M., Gonzalez-Rátiva, M.C. and Orozco-Arroyave, J.R., "Automatic detection of Voice Onset Time in voiceless plosives using gated recurrent units", Digit. Signal Process., 104, 102779, 2020.

[4] Frid, A. and Lavner, Y., "Spectral and textural features for automatic classification of fricatives", XXII Annual Pacific Voice Conference (PVC), 1-4, 2014.

[5] Leemann, A., Kolly, M., Werlen, I., Britain, D. and Studer-Joho, D., "The diffusion of /l/-vocalization in Swiss German", Language Variation and Change, 26(2): 191-218, 2014.

[6] Jongkind, A. P. and van Reenen, P., "The vocalization of /l/ in standard Dutch", in A. Timuska [Ed], Proceedings of the IVth International Conference of Dialectologists and Geolinguists, University of Latvia, Riga, 28.7.2003. Riga: Latvian Language Institute, 1-6, 2007.

[7] Callou, D. and Leite, Y., "Iniciação à Fonética e a Fonologia", Editira Zahar, Rio de Janeiro, 1990.

[8] Hardcastle, W. and Barry, W., "Articulatory and perceptual factors in /l/ vocalizations in English", Journal of the International Phonetic Association, 15:3-17, 1985.

[9] Ash, S., "The vocalization of /l/ in Philadelphia", Doctoral dissertation, University of Philadelphia, Pennsylvania, PA, 1982.

[10] Bauer, L., "English in New Zealand", in Burchfield, R. [Ed], The Cambridge History of the English Language. English in Britain and Overseas: Origins and Development, vol. 5. Cambridge University Press, Cambridge, 82-429, 1994.

[11] Horvath, B. and Horvath, R. J., "The geolinguistics of /l/-vocalization in Australia and New Zealand", Journal of Sociolinguistics, 6(3): 319-346, 2002.
[12] Borowsky, T. and Horvath, B., "L-vocalization in Australian English", in F. Hinskens, R. van Hout, and W. L. Wetzels [Eds], Variation, change and phonological theory, 101-123. Amsterdam: John Benjamins, 1997.
[13] Loakes, D., "An Investigation of the /el/-/ael/ Merger in Australian English: A Pilot Study on Production and Perception in South-West Victoria", Australian Journal of Linguistics, 34: 436-452, 2014.
[14] Hall-Lew, L. & Fix, S., "Perceptual coding reliability of (L)-vocalization in casual speech data", Lingua, 122: 794-809, 2012.
[15] Lin, S. and Demuth, K., "Children's Acquisition of English Onset and Coda /l/: Articulatory Evidence", Journal of Speech, Language, and Hearing Research, 58: 13-27, 2015.
[16] Lin. S., Palethorpe, S. and Cox, F., "An ultrasound exploration of Australian English /CVl/ words', ASSTA, Sydney, Australia, 2012.
[17] Turton, D., "Categorical or gradient? An ultrasound investigation of /l/-darkening and vocalization in varieties of English", Journal of the Association for Laboratory Phonology 8(1): 1-31, 2017.
[18] Szalay, T., Benders, T., Cox, F. and Proctor, M., "Lingual configuration of Australian English /l/", Proceedings ICPhS, 2019.
[19] Scobbie, J.M. and Pouplier, M., "The role of syllable structure in external sandhi: an EPG study of vocalization and retraction in word-final English /l/", Journal of Phonetics, 38: 240-259, 2010.
[20] Yuan, J. & Liberman, M., "Investigating /l/ variation in English through forced alignment", Proceedings Interspeech, 2215-2218, 2009.
[21] Villarreal, D., Clark, L.S., Hay, J. and Watson, K., "From categories to gradience: Auto-coding sociophonetic variation with random forests", Laboratory Phonology, 11(1): 1-31, 2020.
[22] Malakar, M., and Keskar, R.B., "Progress of machine learning based automatic phoneme recognition and its prospect", Speech Commun., 135: 37-53, 2021.
[23] Sloetjes, H. and Wittenburg, P., "Annotation by category – ELAN and ISO DCR", in Proceedings of the 6th International Conference on Language Resources and Evaluation, 2008.
[24] Fromont, R. and Hay, J., "LaBB-CAT: An annotation store", University of Otago, Dunedin, New Zealand: Australasian Language Technology Workshop (ALTA), 4-6 Dec 2012, in Proceedings, 10: 113-117, 2012.
[25] Baayen, R. H., Piepenbrock, R. and Gulikers, L., "CELEX2 LDC96L14. Web Download", Philadelphia: Linguistic Data Consortium, 1995.
[26] Young, S., Evermann, G., Gales, M. et al., "The HTK Book (for HTK Version 3.4)", Cambridge University Engineering Department, 2006.
[27] Boersma, P. and Weenink, D., "Praat: doing phonetics by computer [Computer program]". Version 6.2.19, retrieved 12 September 2022 from http://www.praat.org/.
[28] Durian, D., "The vocalization of /l/ in urban blue collar Columbus, OH African American Vernacular English: a quantitative sociophonetic analysis", OSU Working Papers in Linguistics 58: 30-51, 2008.
[29] Stuart-Smith, J., Lennon, R., Macdonald, R., Robertson, D., Sóskuthy, M., José, B. and Evers, L., "A dynamic acoustic view of real-time change in word-final liquids in spontaneous Glaswegian", 18th ICPhS Proceedings, Glasgow, UK, 2015.
[30] Plug, L. and Ogden, R., "A parametric approach to the phonetics of postvocalic /r/ in Dutch", Phonetica, 60: 159-86, 2003.
[31] R Development Core Team, "R: A language and environment for statistical computing", R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org, 2021.
[32] Kursa, M.B. and Rudnicki, W.R., "Feature Selection with the Boruta Package", Journal of Statistical Software, 36(11): 1–13, 2010.
[33] Stoppiglia H, Dreyfus G, Dubois R. and Oussar, Y., "Ranking a Random Feature for Variable and Feature Selection", Journal of Machine Learning Research, 3: 1399-1414, 2003.
[34] Kohavi, R. and John, G.H., "Wrappers for Feature Subset Selection", Artificial Intelligence, 97: 273-324, 1997.
[35] Breiman, L., "Random Forests", Machine Learning, 45: 5-32, 2001.
[36] Iskarous, K., "The encoding of vowel features in Mel-Frequency Cepstral Coefficients", 2018, downloaded 10.7.22 from https://www.aisv.it/StudiAISV/2018/vol_4/001_Iskarous.pdf
[37] Labov, W., "A sociolinguistic perspective on sociophonetic research", Journal of Phonetics, 34: 500-515, 2006.

Vowel merger in Australian English lateral-final rimes: /æɔ-æ/

Tünde Szalay[1], Titia Benders[2,3], Felicity Cox[2], Michael Proctor[2]

[1] The University of Sydney, Sydney, Australia, [2]Macquarie University, Sydney, Australia,
[3] The University of Amsterdam, Amsterdam, Netherlands

tuende.szalay@sydney.edu.au

Abstract

Pre-lateral contrast reduction between Australian English /æɔ-æ/ (*howl-Hal*) compared to other environments might indicate an ongoing merger. Our apparent-time study explores this merger. Spectral and temporal characteristics of /æɔ-æ/ produced in pre-lateral and pre-obstruent contexts by 19 older and 15 younger speakers were compared. Acoustic vowel similarity was captured using random forest classification and hierarchical cluster analysis of dynamic formant properties and duration values. Consistent with a pre-lateral merger, younger speakers showed reduced pre-lateral vowel contrast than older speakers, and young male speakers produced /æɔl-æl/ similarly to pre-obstruent /æɔ/. Pre-lateral /æɔ-æ/ merger is attributed to younger speakers' changing *F*2 trajectories.

Index Terms: vowel change, Australian English, pre-lateral vowels, change by coarticulation

1. Introduction

Contrast reduction caused by systematic and directional coarticulatory variation is often implicated in the initiation of sound change, yet not all coarticulatory variation leads to sound change [1, 2, 3, 4]. In the Interactive-Phonetic (IP) model, sound change may be initiated when highly coarticulated realisations of one phoneme become acoustically similar to another phoneme, while other realisations remain distinct [4]. That is, sound change of this type has two important features: one phoneme must shift in the acoustic space according to its coarticulatory context, and another phoneme must already occupy the acoustic space that the coarticulated allophone is moving into. As listeners and speakers interact, coarticulated realisations are incorporated into listeners' representation, shifting them closer to the second phoneme, and potentially leading to a merger [4]. Such a merger is signalled by failed compensation for coarticulation in the IP model [4].

The Australian English (AusE) vowel pairs /iː-ɪ, ʉː-ʊ, æɔ-æ, əʉ-ɔ/ (e.g. *feel-fill, fool-full, howl-Hal, dole-doll*) may satisfy the necessary conditions of vowel change through coarticulation: members of the pairs show acoustic and perceptual contrast reduction in the pre-lateral position, while their pre-obstruent allophones remain distinct [5, 6, 7]. Decreased spectral contrast between pre-lateral /æɔ-æ/ is attributed to the *F*2 trajectory of /æl/ (Hal) becoming similar to that of /æɔl/ (*howl*) [6]. As the tongue transitions from the front vowel to dorsal dark /l/ in /æl/, a falling *F*2 transition is created, similar to that of the front-back transition in the diphthong /æɔ/ [6]. The spectral contrast reduction corresponds to perceptual contrast reduction as listeners are likely to confuse members of the /æɔ-æ/ vowel pair in the /l/ context [7]. In the pre-obstruent context, spectral and perceptual vowel contrast is preserved due to the maintenance of the high *F*2 of /æ/ [6, 7].

Contrast reduction between pre-lateral /æɔ-æ/ may be consistent with a contextual vowel merger in the IP model of sound change. Vowel-lateral coarticulation creates vowel realisations that potentially lead to overlapping formant trajectories and durations in separate phonemes. However, apparent-time studies have not examined pre-lateral merges in AusE, only in other varieties of English, such as American and British English [8, 9, 10]. Therefore, we examine if there is a pre-lateral vowel change and merger between /æɔ-æ/ (*howl-Hal*) in AusE. We hypothesised that younger speakers would (1) show smaller contrast between pre-lateral allophones of the members of the vowel pair /æɔ-æ/ than older speakers; (2) shift their production of /æl/ towards /æɔl/; and (3) younger and older speakers would preserve /æɔ-æ/ contrast in the pre-obstruent environment.

2. Methods

2.1. Speakers

Data were extracted from AusTalk, an AusE speech corpus recorded between 2011 to 2015 [11]. Speech recordings of 15 younger (F = 8, M = 7, ages = 20 – 29, mean = 23.5) and 19 older (F = 9, M = 10. ages = 51 – 80, mean = 60.5) native speakers of AusE were selected from the database. Speakers were born and educated in the Greater Sydney Metro Region with at least one of their parents born in Australia. The speakers did not report any reading, speaking, or hearing difficulties.

2.2. Material and procedure

The two stressed vowels /æɔ-æ/ were produced in two monosyllabic paradigms, /hVd/ and /hVl/ (*howd-had, howl-Hal*), in a single-word production task. Speakers read 322 isolated words, including the four target words, as they were presented orthographically on a computer monitor in a random order. The task was recorded on three separate occasions, each using a different order of words. Each speaker produced up to three repetitions of each lexical item; the number of repetitions differs between participants, as not all participants attended all three sessions.

2.3. Phonetic analysis

400 tokens were analysed (4 target words × 34 speakers × 3 maximal repetitions - 8 missing repetitions). Segment boundaries were automatically located using the MAUS forced aligner with the AusE grapheme-to-phoneme converter [13, 14, 15], and manually corrected in a Praat interface [16]. The vowel onset was determined on the basis of voicing onset and sudden increase in amplitude. Vowel offset in the /d/ context was determined on the basis of amplitude drop. Rime offset in the /l/ context was determined on the basis of voicing offset. Because there is no discernible boundary between the vowel and the fol-

lowing /l/ in /hVl/ words, the entire /hVl/ rime was analysed instead of selecting an arbitrary boundary in the vowel-lateral transition (Fig. 1). Automatic segmentation errors were corrected only when the boundary was judged to be misplaced by more than 20 ms [17]. Boundary correction was carried out by the first author and a phonetically trained research assistant with 15% of the data cross-marked by both. Boundary agreement, with a 20 ms agreement threshold was 99% for vowel onsets and 97% for vowel offset. /l/ offset boundaries were re-checked and corrected if necessary by the first author as agreement was 60%.

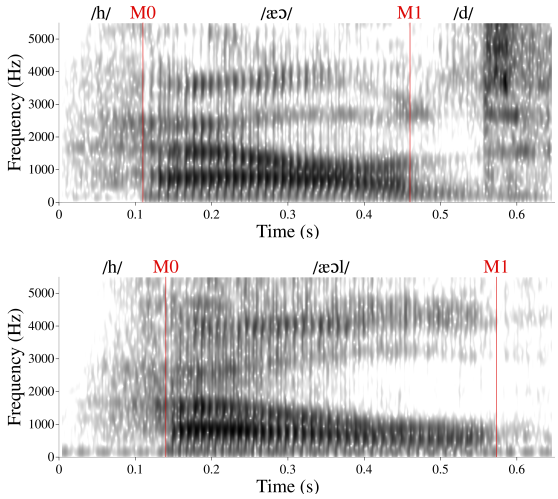


Figure 1: *Vowel (top) and rime (bottom) onset and offset.*

Formant trajectories in the pre-/d/ vowels and in the lateral-final rimes were extracted automatically and corrected manually in Praat [16]. Formant frequencies were estimated at every 5 ms throughout a 25 ms formant analysis window using a 50 ms Gaussian window with 75% overlap and with 50 dB dynamic range and a pre-emphasis filter increasing spectral slope above 100 Hz by 6 dB/octave. To optimise formant settings for each speaker, five to six formants were tracked up to 4500 ceiling for speakers who produced comparatively lower $F2$ and $F3$ or up to a maximum frequency of 7000 Hz for speakers who produced a comparatively higher $F2$ or $F3$ trajectory. Male speakers were typically analysed with lower and female speakers with higher formant ceiling. Formant trajectories were manually corrected using a Praat-based interface that superimposed formant estimates over a broadband spectrogram calculated over 5 ms windows with 40% overlap, allowing for corrections of estimates that did not align with the visible formants. Manual correction was carried out by the first author and a phonetically trained research assistant. After hand-correction, $F1$–$F3$ trajectories for every word were visually inspected by the first author; values 1.5 times above or below the interquartile range for each formant in each vowel × coda × age × gender group were rechecked by the first author.

Discrete cosine transformations (DCT) were used to model formant change over time using the first three DCT coefficients [18, 19, 20]. The 0^{th} coefficient represents the mean of a formant trajectory multiplied by $\sqrt{2}$; the 1^{st} coefficient represents the direction and magnitude of the curve of the trajectory; the 2^{nd} coefficient represents the trajectory's curvature. Each token was represented parametrically by a total of 9 DCT coefficients (3 formants × 3 coefficients).

2.4. Statistical analysis

To test spectral similarity, we trained two random forest classification models to learn 2 (vowels) × 2 (coda) × 2 (age) = 8 categories for male and 8 categories for female speakers based on the DCT coefficients, duration values, and group labels using 75% of data produced by each gender [21, 23]. The training phase returned an out-of-bag error rate. A low out-of-bag error rate indicates that the algorithm was successful at learning the categories [22]. The remaining 25% of the tokens were used to test the classifier, by grouping unlabelled values based just on DCT coefficients and duration values. The testing phase returns two confusion matrices, separately for each gender. The confusion matrices were then fed into an agglomerative hierarchical cluster analysis using Ward's method [24] to measure between-group similarity based on the confusability rates. The results of hierarchical cluster analysis are represented on a dendrogram: elements that are clustered together are similar to each other, and the lower the cluster is split from the other elements, the higher the spectral similarity between the members of the cluster. That is, the location of nodes can be used for comparing between-cluster similarity. Approximately Unbiased p-value for each multi-element cluster were extracted by repeating the hierarchical cluster analysis on the same confusion matrices using multiscale bootstrap sampling [25, 26]. Approximately Unbiased p-value expresses the frequency with which a cluster appears in bootstrapping; the significant threshold is 95% or above. (For more details on the statistical analysis, see [6]).

Durational contrast reduction in lateral-final rimes compared to pre-/d/ vowels was further tested using generalised linear mixed models (GLMs) [27]. A GLM model was built with the dependent variable Duration, and the independent variables Coda, Vowel, Age, and Gender. Independent variables were contrast coded, giving pre-/d/ /æɔ/ produced by older female speakers as a baseline. Speaker was added as random intercept with Coda for random slope; random slope for Age and Gender was not added as the study had a between participant design. Convergence was estimated using the BOBYQA (Bound Optimization BY Quadratic Approximation) optimizer and an increased number of maximum iterations [28]. p-values were calculated using Satterthwaite's degrees of freedom method [29]. All statistical analyses were carried out in R [31].

3. Results

Two random forest classification models were trained on DCT coefficients, duration values, and group labels using 75% of the data for each genders. Out-of-bag error rate in the testing phase is 36.67% for male speakers and 40.67% for female speaker, indicating that DCT coefficients and duration values can classify vowels with comparable accuracy for both genders. Although out-of-bag error rates are high, they are in line with rates observed for classifying /l/-final rimes in AusE [6].

We used hierarchical cluster analysis to test whether confusion rates are statistically significant. Male speakers show five significantly frequently occurring clusters (Fig. 2). The two clusters of /d/-final rimes (100% frequency for both) are split by vowel and merged by age, indicating that the vowels /æ/ and /æɔ/ are produced similarly between age groups and differently from each other in the /d/ context. The clusters of /l/-final rimes are split by age (97% for older, and 100% for younger speakers) and merged by vowel, indicating that /l/-final rimes differ between the age groups but are similar between the vowels /æɔ/ and /æ/. In addition, there is a supercluster consisting of all male

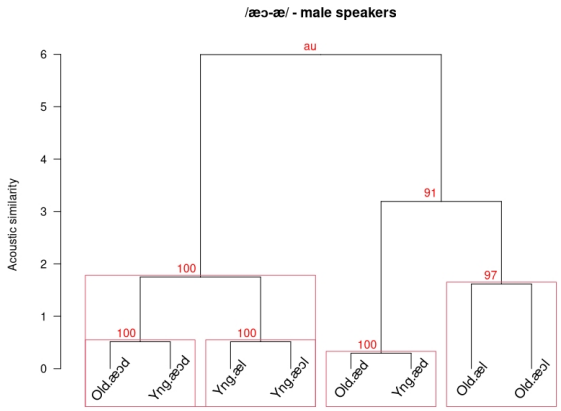


Figure 2: *Acoustic similarity in male speakers. Y-axis: similarity increases as the number decreases; similarity measured in arbitrary units. Au: approximately unbiased p-values.*

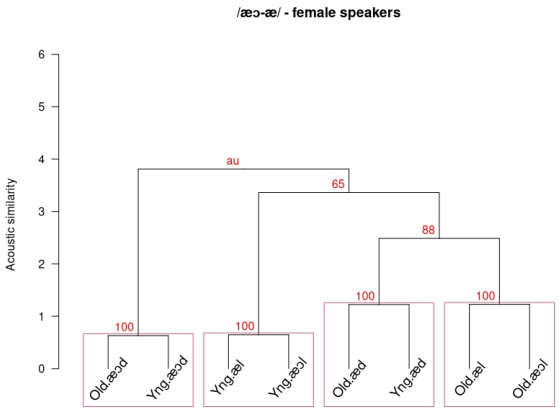


Figure 3: *Acoustic similarity in female speakers. Y-axis: similarity increases as the number decreases; similarity measured in arbitrary units. Au: approximately unbiased p-values.*

speakers' pre-/d/ diphthong and the younger speakers' /l/-final rimes (100% frequency), indicating that lateral-final /æɔ-æ/ shift toward the diphthong for young male speakers.

Female speakers show four clusters that occur with significant frequency (100% for each) (Fig. 3). The two /d/-final clusters are split by vowel and merged by age, indicating that the vowels /æ/ and /æɔ/ are produced similarly between age groups and differently from each other in the /d/ context. The two /l/-final clusters are split by age and merged by vowel, indicating that /l/-final rimes differ between the age groups but are similar between the vowels /æɔ/ and /æ/.

Younger speakers' pre-lateral clusters of /æɔ-æ/ branch at a lower point (below one) compared to older speakers (above one) for both genders, indicating higher higher similarity between members of the vowel pair in pre-lateral position (Figs. 2 – 3). As /l/-final rimes are inherently longer than pre-/d/ vowels, it is possible that /l/-final rimes formed separate clusters from pre-/d/ vowels only due to their different durational values. Therefore, we repeated random forest and hierarchical cluster analyses for both genders without duration values. For male speakers, /l/-final rimes formed separate clusters from pre-/d/ vowels for older speakers, but younger speakers' /l/-final rimes clustered with pre-/d/ /æɔ/ with and without duration values. For female speakers, /l/-final rimes formed separate clusters from pre-/d/ vowels for older and younger speakers with and without duration values. That is, the cluster of /l/-final rimes and the clusters of pre-/d/ vowels in Figs. 2–3 are not formed on the basis of duration values alone for any of the genders.

Our GLM shows that duration was not affected significantly by the main effect of Age and Gender, indicating no significant difference in /æɔ/ duration between the age groups and genders. The short vowel /æ/ was significantly shorter than the long vowel /æɔ/ ($\beta = -127.93, t_{9.84} = -13, p < 0.0001$). Coda /l/ resulted in a significant increase in duration for /æɔl/ compared to the vowel in the /d/-context ($\beta = 119.96, t_{15.81} = 7.59, p < 0.0001$) (Fig. 4).

An interaction between Coda /l/ and Vowel indicates that the rime containing the short vowel lengthens more in the /l/ context compared to the rime containing the diphthong ($\beta = 45.85, t_{13.92} = 3.29, p < 0.0011$). The remaining interactions, including the ones with Age, were not significant (Fig. 4).

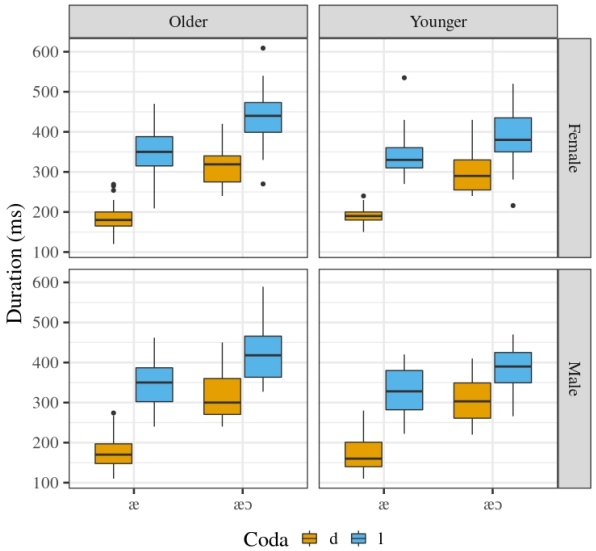


Figure 4: *Rime- and vowel duration.*

4. Discussion

Our first hypothesis stated that younger speakers would show smaller contrast between pre-lateral allophones of the members of the vowel pair /æɔ-æ/ compared to older speakers. Although this vowel contrast is reduced in pre-lateral environments compared to pre-obstruant contexts for both older and younger speakers, in line with our hypothesis, younger speakers show smaller contrast compared to older speakers. Smaller acoustic contrast between pre-lateral rimes for younger speakers compared to older speakers is consistent with an ongoing pre-lateral vowel merger.

The increased contrast reduction between pre-lateral /æɔ-æ/ for younger speakers might be driven by contrast reduction between the /æ/-/l/ transition and the second target of the diphthong /æɔ/ (120 - 300 ms, between vertical lines in the right panel of Fig. 5). Young speakers produce the /æ/-/l/ transition with a less steep $F2$ drop compared to older speakers, shifting their /æ/-/l/ transition closer to the second target of the diph-

thong /æɔ/. In contrast, older speakers preserve the *F*2 contrast between between the /æ/-/l/ transition and the second target of the diphthong /æɔ/ as they produce the /æ/-/l/ transition with a lower *F*2 compared to the second target of the diphthong. To explore at which point in time contrast reduction takes place in the rime, future research is required to examine formant trajectories using Generalised Additive Mixed Models.

Younger speakers appear to reduce duration contrast between lateral-final /æɔl/ and /æl/ compared to older speakers (Fig. 5), which might contribute to their increased contrast reduction. To address the role of duration contrast, future research is required to separate the vowel from the /l/.

Our second hypothesis stated that contrast reduction would be shown in the pre-lateral monophthong shifting toward the pre-lateral diphthong. In line with our hypothesis, contrast reduction between /æl/ and /æɔl/ is driven by the *F*2 transition from the monophthong /æ/ towards coda /l/: pre-obstruent /æ/ has a steady high *F*2 throughout the vowel (0-250 ms, left panel of Fig. 5), while pre-lateral /æ/ shows an *F*2 decline (from 25-50 ms onward, right panel of Fig. 5). As the tongue moves from the front vowel target to the dorsal target of dark coda /l/, it creates a back vowel-like transition, making /æl/ similar to /æɔ/ [32].

In addition, the vowel /æ/, when coarticulated with /l/, shows more similarities with pre-/l/ and pre-/d/ /æɔ/ for younger male speakers compared to older male speakers (Figs. 2 and 6). Increased similarity between lateral-final rimes and the pre-obstruent diphthong might be driven by the rising *F*2 at the end of the lateral-final rimes in young speakers' production (Figs. 5-6). This *F*2 increase is not consistent with the small *F*1-*F*2 of dark /l/ or with the low *F*2 of dark or vocalised /l/ [33, 34]. An increased *F*2 is consistent with young speakers vocalising *less* than older speakers (contrary to [35]). Alternatively, as the end of the analysis window was defined by the end of voicing, it may be caused by the tongue moving toward a neutral rest position. An articulatory study is required to address /l/-vocalisation.

Female speakers do not show acoustic similarities between lateral-final rimes and the pre-obstruent diphthong, despite showing similar *F*2 transitions from /æ/ towards coda /l/ as from [æ] to [ɔ] (Fig. 5). This difference might arise from timing differences: the duration of /l/-final rimes is longer than that of the pre-/d/ diphthong for female speakers (Figs. 4-5). For young male speakers, duration of /l/-final /æɔl/ and /æl/ are not too dissimilar from pre-/d/ /æɔ/. Coda /l/ lengthens rimes containing the short vowel more, indicating duration contrast reduction with no difference between genders; however, hierarchical cluster analysis might have been sensitive for small differences in duration contrast produced by male and female speakers.

Our third hypothesis stated that pre-/d/ allophones of the members of the vowel pair /æɔ-æ/ would remain distinct. We found no evidence of pre-obstruent contrast reduction for /æɔ/ and /æ/, as the vowels never clustered together in the pre-/d/ context due to their spectral and durational differences. Therefore, contrast reduction between /æɔl - æl/ is attributed to coarticulation with coda /l/ rather than across the board vowel contrast reduction. Across the board vowel change can be observed as younger speakers produce /æ/ with a higher *F*1 and lower *F*2 compared to older speakers (Fig. 5). However, in the pre-obstruent context, younger speakers' /æ/ shows spectral similarities to older speakers' /æ/, and younger speakers' /æɔ/ shows spectral similarities to older speakers' /æɔ/. Thus, as predicted by the IP model of sound change [4], AusE /æɔ-æ/ show an ongoing pre-lateral vowel merger caused by the coarticulatory influence of /l/, as the pre-lateral allophone of /æ/ moves through a similar acoustic space as /æɔ/, while the pre-obstruent allophones remain distinct. The shift of /æl/ to /æɔl/ in production is consistent with /æl/ being more likely to be misperceived as /æɔl/ than /æɔl/ as /æl/ [7]. Future research is required on the effect of speakers' and potentially listeners' age and gender on pre-lateral vowel contrast reduction.

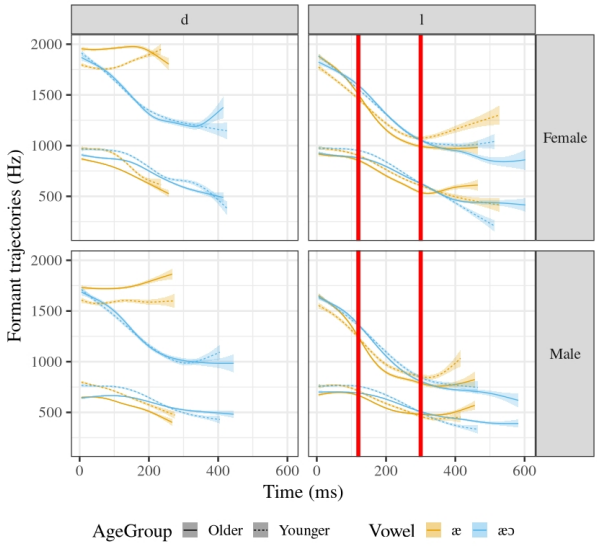


Figure 5: *F1-F2 trajectories. Vertical lines: areas of interest highlighting potential F2 contrast reduction in young speakers.*

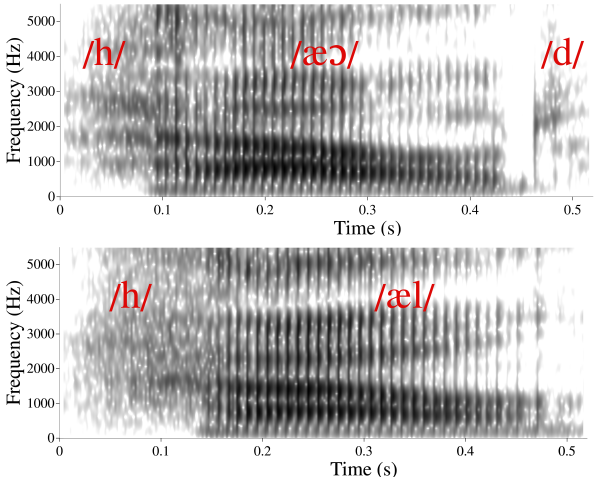


Figure 6: *Young male speaker's production. Top: Pre-obstruent /æɔ/ (howd) Bottom: Pre-lateral /æ/ (Hal). Colour on-line.*

5. Conclusions

Pre-lateral vowel merger between members of the vowel pair /æɔ-æ/ is shown, as younger speakers produce members of the pairs with smaller spectral contrast compared to older speakers. Pre-lateral /æ/ shifts toward /æɔ/ due to the coarticulatory influence of /l/. Male speakers' lateral-final rimes shift toward pre-obstruent /æɔ/, while female speakers only reduce contrast between the two lateral-final rimes. Decreased contrast may be carried by changes in young speakers' production of the /æ/-/l/ transition: while older speakers maintain a contrast between /æ/-/l/ transition and the second target of /æɔ/, younger speakers æ/-/l/ transition shifts towards [ɔ].

6. Acknowledgements

We thank Ms. Aimee Death for her help in phonetic data analysis. This research was supported in part by Australian Research Council DE150100318 and Australian Research Council Future Fellowship Grant FT180100462.

7. References

[1] Ohala, J. J., "Sound change is drawn from a pool of synchronic variation", in L. E. Breivik, and E. H. Jahr [Eds], Language change: Contributions to the study of its causes, 173–198, Mouton de Gruyter, 1989.

[2] Blevins, J., "A theoretical synopsis of Evolutionary Phonology", Theoretical Linguistics, 117–165, 2006.

[3] Garrett, A. and Johnson, K. "Phonetic bias in sound change", in A. C. L. Yu [Ed], Origins of Sound Change, 51–97, 2013.

[4] Harrington, J., Kleber, F., Reubold, U., Schiel, F. and Stevens, M., "Linking Cognitive and Social Aspects of Sound Change Using Agent-Based Modeling", Topics in cognitive science, 10(4):707–728, 2018.

[5] Palethorpe, S., and Cox, F. "Vowel modification in pre-lateral environments", International Seminar on Speech Production, 2003.

[6] Szalay, T., Benders, T., Palethorpe, S., Cox, F. and Proctor, M., "Spectral contrast reduction in Australian English /l/-final rimes", J. Acoust. Soc. Am., 149(2):1183–1197, 2021.

[7] Szalay, T., Benders, T., Cox,F. and Proctor, M., "Perceptual vowel contrast reduction in Australian English /l/-final rimes", Laboratory Phonology 12(1), 2021.

[8] Kleber, F., Harrington, J. and Reubold, U.,"The relationship between the perception and production of coarticulation during a sound change in progress", Language and Speech, 55(3):383–405, 2012.

[9] Arnold, L., "Multiple mergers: Production and perception of three pre-/l/ mergers in Youngstown, Ohio", University of Pennsylvania Working Papers in Linguistics 21(2), 2015.

[10] Strycharczuk, P. and Scobbie, J. M, "Fronting of Southern British English high-back vowels in articulation and acoustics", J. Acoust. Soc. Am., 142(1):322–331, 2017.

[11] Burnham, D., Estival, D., Fazio, S., Viethen, J. Cox, J., Dale, R., Cassidy, S., Epps, J., Togneri, R., Wagner, M., Kinoshita, Y., Göcke, R., Arciuli, J., Onslow. M., Lewis, T., Butcher, A. and Hajek, J., "Building an audio-visual corpus of Australian English: large corpus collection with an economical portable and replicable Black Box", Proc Interspeech, 841–844, 2011.

[12] Kisler, T., Reichel, U. D. and Schiel, F. "Multilingual processing of speech via web services"

[13] Reichel, U. D. "PermA and Balloon: Tools for string alignment and text processing", Proc Interspeech, 2012.

[14] Schiel, F. "Automatic Phonetic Transcription of Non-Prompted Speech", Proc ICPhS 607-610, 1999.

[15] Schiel, F. "A Statistical Model for Predicting Pronunciation", Proc ICPhS, 2015.

[16] Boersma, P. and Weenink, D., "Praat: doing phonetics by computer". Version 6.1.39, retrieved 8 February 2021 from www.praat.org

[17] Cosi, P., Falavigna, D. and Omologo. M., "A preliminary statistical evaluation of manual and automatic segmentation discrepancies", Proc. Eurospeech, 693–696, 1991.

[18] Harrington, J. and Cassidy, S., "Dynamic and target theories of vowel classification: Evidence from monophthongs and diphthongs in Australian English", Language and Speech 37(4):357–373, 1994.

[19] Watson, C. I. and Harrington, J., "Acoustic evidence for dynamic formant trajectories in Australian English vowels", J. Acoust. Soc. Am., 106(1):458–468, 1999.

[20] 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.

[21] Burger, S. V., Introduction to Machine Learning with R: Rigorous Mathematical Analysis, O'Reilly Media, Inc., 2018.

[22] Liaw, A. and Wiener, M., "Classification and Regression by randomForest", R News 2(3), 2002.

[23] Breiman, L. "Manual on setting up, using, and understanding random forests v3. 1", Statistics Department University of California Berkeley, CA, USA 1(58):3-42, 2002.

[24] Ward, J., H., "Hierarchical Grouping to Optimize an Objective Function", J. Am. Stat. Assoc., 301(58):236–244, 1963.

[25] Ryota S., Yoshikazu T., and Hidetoshi S., "pvclust: Hierarchical Clustering with P-Values via Multiscale Bootstrap Resampling", R package version 2.2-0, 2019.

[26] Efron, B., Halloran, E. and Holmes, S., "Bootstrap confidence levels for phylogenetic trees", Proc National Academy of Sciences 93(23):13429–13434, 1996.

[27] Bates, D., Mächler, M., Bolker, B., and Walker, S. "Fitting Linear Mixed-Effects Models Using lme4", J Stat Softw 67(1):1–48, 2015.

[28] Powell, M. J. D. The BOBYQA algorithm for bound constrained optimization without derivatives. Cambridge NA Report NA2009/06, University of Cambridge, Cambridge, 26–46, 2009.

[29] Kuznetsova, A., Brockhoff, P. B. and Christensen, R. H. B., "lmerTest Package: Tests in Linear Mixed Effects Models", J Stat Softw 82(13):1–26, 2017.

[30] Lenth, R.V., "emmeans: Estimated Marginal Means, aka Least-Squares Means", 2021.

[31] R Core Team, "R: A Language and Environment for Statistical Computing", R Foundation for Statistical Computing, 2021.

[32] Gick, B., Kang, M. A. and Whalen, D. H., "MRI evidence for commonality in the post-oral articulations of English vowels and liquids", J. Phon 30(3): 357–371, 2002.

[33] Sproat, R. and Fujimura, O., "Allophonic variation in English /l/ and its implications for phonetic implementation", J. Phon., 21(3):291–311, 1999.

[34] Strycharczuk, P. and Scobbie, J. M., "Gradual or abrupt? The phonetic path to morphologisation", Phonetica 59: 76–91, 2016.

[35] Horvath, B. M. and Horvath, R. J., "The geolinguistics of /l/ vocalisation in Australia and New Zealand," Journal of Sociolinguistics 6(3):319–346, 2002.

OBISHI: Objective Binaural Intelligibility Score for the Hearing Impaired

Candy Olivia Mawalim, Benita Angela Titalim, Masashi Unoki, and Shogo Okada

Japan Advanced Institute of Science and Technology,
1-1 Asahidai, Nomi, Ishikawa 923–1292 Japan
{candylim, s2110104, okada-s, unoki}@jaist.ac.jp

Abstract

Speech intelligibility prediction for both normal hearing and hearing impairment is very important for hearing aid development. The Clarity Prediction Challenge 2022 (CPC1) was initiated to evaluate the speech intelligibility of speech signals produced by hearing aid systems. Modified binaural short-time objective intelligibility (MBSTOI) and hearing aid speech prediction index (HASPI) were introduced in the CPC1 to understand the basis of speech intelligibility prediction. This paper proposes a method to predict speech intelligibility scores, namely OBISHI. OBISHI is an intrusive (non-blind) objective measurement that receives binaural speech input and considers the hearing-impaired characteristics. In addition, a pre-trained automatic speech recognition (ASR) system was also utilized to infer the difficulty of utterances regardless of the hearing loss condition. We also integrated the hearing loss model by the Cambridge auditory group and the Gammatone Filterbank-based prediction model. The total evaluation was conducted by comparing the predicted intelligibility score of the baseline MBSTOI and HASPI with the actual correctness of listening tests. In general, the results showed that the proposed method, OBISHI, outperformed the baseline MBSTOI and HASPI (improved approximately 10% classification accuracy in terms of F1 score).

Index Terms: hearing impaired, speech intelligibility, binaural hearing, hearing aids, hearing loss model

1. Introduction

Hearing aids are a technology that contributes to assisting sensorineural hearing loss. The hearing loss phenomenon can be explained in several ways. First, the auditory threshold is lifted above 0 dB or above the auditory threshold in normal hearing (NH). Second, the contribution of hair cells in inner ear damage to the signal compression and auditory threshold is shifted to the higher range [1, 2]. These factors describe how the damage in the inner ear and the noise level affect speech perception, and hearing aids should compensate for the loss. Speech processing is needed in hearing aids to enhance speech quality and intelligibility, especially in noise and reverberation.

One of the important evaluations for hearing aids is the speech intelligibility metrics. Speech intelligibility often refers to how accurately speech is understood or the percentage of the number of words the listener correctly identifies [3, 4]. The hearing aid speech prediction index (HASPI) by Kates and Arehart [5, 6] is often considered in developing hearing aids as an objective speech intelligibility index. The HASPI model includes a comparison of the temporal amplitude envelope (TAE) and temporal fine structure (TFS) that makes the prediction accuracy in both NH and hearing-impaired (HI) processing improved [5]. Unfortunately, the HASPI model has several drawbacks; that is, evaluation is limited to the conditions provided in the training data, handles monaural listening, only considers the audiogram for the listener's hearing characteristics, and is invalid for tonal languages.

Another alternative to measuring objective speech intelligibility is the modified binaural short-time objective intelligibility (MBSTOI) [7]. This model was developed based on the STOI metric [8] and is an extended model of discrete binaural STOI (DBSTOI) [9]. The MBSTOI generates more accurate predictions than the DBSTOI because it overcomes the tendency of overestimation when the interferers are spatially distributed. However, this model utilized a hearing loss model [10] to approximate the HI auditory thresholds by adding internal noise and by attenuating the signals. Thus, the baseline model is sensitive to the level of the processed signal.

This study proposes an objective binaural intelligibility score for the hearing impaired (OBISHI) to improve the speech intelligibility performance of existing methods. For instance, unlike the HASPI model, the proposed method handles binaural listening. Additionally, the proposed method considers not only the listener's audiogram but also other HI characteristics, such as the digit-triplet test (DTT) results. The proposed prediction model integrates a pre-trained automatic speech recognition (ASR) system to predict the difficulty of the sentence regardless of the hearing loss conditions, an HI characteristics (HICs) predictor, and an intelligibility model built on a gammatone filterbank.

2. Hearing-Impaired Intelligibility Model

The Clarity Challenge[1] was formed as one part of contributing to the development of hearing aid technology to improve the signal processing in the hearing aids system and to predict the perceived speech in noise (SPIN). One of the main tasks of this challenge is to predict the speech intelligibility of HI listeners when they perceive noisy speech processed by a hearing aid system [4]. It provides audio signals from simulated hearing aids receiving SPIN with the corresponding reference signals & transcript, the HI listeners' characteristics, and the speech intelligibility score as the ground truth obtained from listening tests. The simplified baseline system consists of a hearing loss simulation and binaural speech intelligibility models. However, the configuration of the prediction model can be altered, for example, by combining the hearing loss and speech intelligibility model with a single model. Two HI intelligibility models are also introduced in CPC1: HASPI and the baseline MBSTOI models.

[1] http://claritychallenge.org/

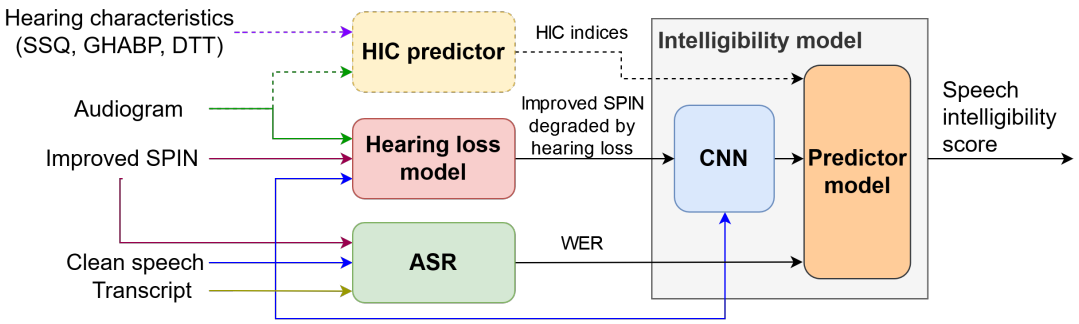


Figure 1: Block diagram of our proposed method. The dashed components were optionally included in the close-set prediction model. The HIC predictor is excluded in the open-set prediction model due to the unknown hearing characteristics of the unseen listener.

2.1. Hearing Aid Speech Prediction Index

HASPI is an index to predict speech intelligibility for NH and HI listeners. The model consists of analysis by a fourth-order gammatone filterbank (GTFB) with bandwidth changes for HI by a control filterbank. The control filterbank performs the processing of loudness recruitment and the signal intensity effect [11, 12] based on outer-hair cell (OHC). Besides, the damage to the inner hair cell (IHC), as a transducer that transmits mechanical force to an electrical signal to the brain, reduces transduction efficiency and results in additional attenuation [2]. The attenuation is represented by IHC firing-rate adaptation [13] in this model. The HASPI measurement generates cepstral coefficient sequences from the envelope output of the auditory periphery. The cepstral coefficient sequences then pass through modulation filters to give filtered sequences. The normalized cross-covariance is calculated and averaged across the basis function to produce an averaged vector of covariances. Finally, the speech intelligibility score is mapped using neural networks.

2.2. Modified Binaural Short-Time Objective Intelligibility

The baseline of the speech intelligibility prediction model for the Clarity Prediction Challenge 2022 [4] was the MBSTOI [14]. As the name suggests, MBSTOI is constructed based on the STOI metric that takes advantage of non-linear processing with no separate noise access requirement. The model also takes advantage of the first extended version, deterministic binaural STOI (DBSTOI), which covers the binaural processing and fluctuating interferers [7]. The intermediate correlation coefficient of the degraded and reference signals is calculated similarly to that in the STOI measurement. Then, the intelligibility score is defined as the average of the intermediate correlation across time and ⅓ octave frequency bands.

The baseline MBSTOI is constructed to predict the SPIN in broader types of noise, spatially different interferers, linear and non-linear processing, and reverberation. It corrects the broadband delay of the ear model due to hearing loss by running the Kronecker delta function. However, the baseline model did not fix the delay after the hearing aid processing. Moreover, a clean reference was not provided for the tests or evaluation set to correct the hearing aid processing delay in the challenge. Another problem with this model is that the MBSTOI is insensitive to the processed signal level. The measurement is based on the cross-correlation method, which may produce a highly intelligible result when the sound falls below auditory thresholds.

3. Proposed Method

The overall process in our method is shown in Fig. 1. The general inputs for both models are clean speech, improved SPIN (the output of the hearing aid system), and the audiogram of the HI listener. The HI characteristics (HIC) of the listener, including the results of speech, spatial, and qualities of hearing scale 12 (SSQ12) [15], the Glasgow hearing aid benefit profile (GHABP) [16], and the digit-triplet test (DTT) [17] were taken into account for inferring the HIC indices.

Our method has four main components: a HIC predictor, ASR, a hearing loss model, and an intelligibility model. The HIC predictor receives the SSQ, GHABP, DTT, and audiogram, resulting in the HIC indices representing each listener's characteristics. The imputation approach handled the missing data in the HIC characteristics, where the mean value is used to fill the missing data of the SSQ and GHABP of the available listeners. Meanwhile, we determined the DTT results using a prediction model by inputting the other listeners' characteristics and audiogram.

The ASR receives the clean speech and the output SPIN as inputs, and it outputs the word error rate (WER) of the predicted sentence of the output SPIN with the predicted sentence of the clean speech as the reference. We utilized a pre-trained ASR system [18] built using a factorized time delay neural network (TDNN-F) [19] that was trained on the LibriSpeech dataset [20]. The purpose of integrating an ASR in our model was to predict the sentence difficulty regardless of the HI condition (the recognition rate for the NH listener). The hearing loss model by the Cambridge auditory group [10] was utilized to estimate the improved SPIN degraded by hearing loss. The model, namely Moore, Stone, Baer, and Glasberg (MSGB) model, comprises simulations of acoustic transformation in the cochlea, spectral smearing and threshold elevation, and loudness recruitment.

Figure 2 shows a block diagram of how to generate the Convolutional Neural Network (CNN) input for our intelligibility model. We consider the speech inputs to be binaural signals. An infinite impulse response (IIR) time-domain GTFB[2] [21] with 32-channels was utilized to analyze the signals from both ears. Subsequently, we extracted the TAEs from the output of each channel in the GTFB analysis. These TAEs were then passed through a CNN (as shown in Fig. 3). The final predictor model consists of two layers of a fully-connected network with a recti-

[2] https://github.com/huynguyenqc/MOSA-Net-Cross-Domain/tree/VoiceMOSChallenge/gammatone_iir

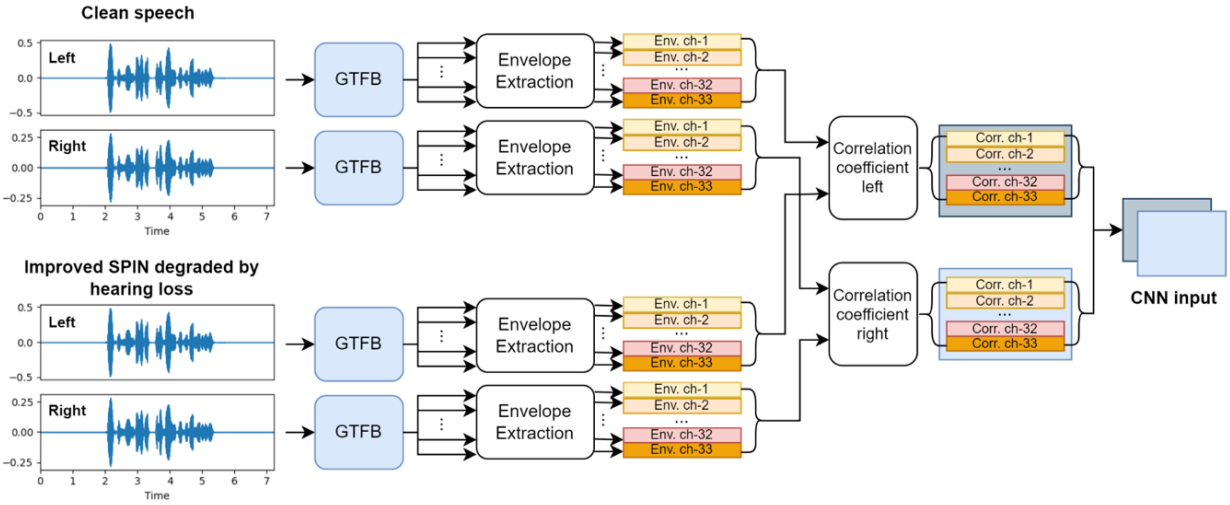


Figure 2: CNN input generation

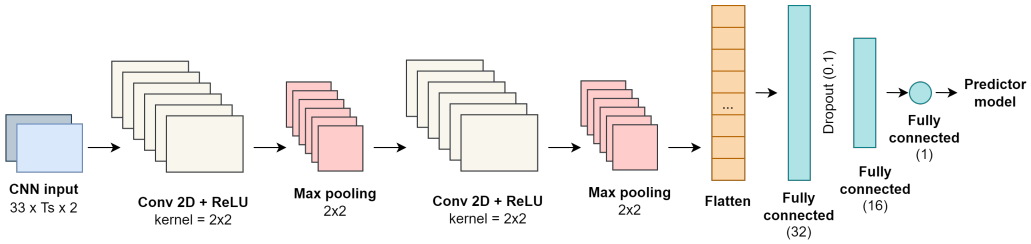


Figure 3: CNN architecture

fied linear unit (ReLU) activation function that receives the output of the CNN layer, the HIC indices, and the WER to predict the speech intelligibility score. We used the Adam optimizer algorithm and the mean-squared error (MSE) loss function in the training process.

4. Evaluation

4.1. Dataset

We utilized the dataset available in the Clarity Prediction Challenge 1 (CPC1)[3] [14]. Generally, it consists of a relatively large number of 44.1-kHz, 32-bit mono or stereo wav files and their corresponding metadata. The wav files are generated scenes, interferers, original target speech spoken by British English speakers, and improved SPINs (the output of SPINs after passing through hearing aid processors). The metadata provides detailed information related to the scenes, listeners, and transcripts. The dataset has six speakers, ten hearing aid processors in the first Clarity Enhancement Challenge [14], and 27 HI listeners. The hearing ability conditions of each listener are also available, including the pure-tone air-conduction audiogram for both ears, the DTT [17] results, and two self-assessment results (i.e., SSQ12 [15] and the GHABP questionnaire [16]). Unfortunately, some of the data of the DTT, SSQ12, and GHABP questionnaire results were missing for several listeners.

[3] https://claritychallenge.github.io/clarity_CPC1_doc/docs/cpc1_data

CPC1 has two tracks: track 1 (close-set) and track 2 (open-set). Each track has a different distribution of training/development and testing sets. The training/development and testing sets for both tracks do not overlap. We split the training/development data of track 1 (4,863 scenes) into 90% for training data and 10% for development data. The testing data of track 1 consists of 2,421 scenes. Track 2 consists of 3,580 training/development scenes and 632 test scenes. We split the training and development data by a leave-one-listener-and-one-system-out approach. This approach results in 2,933 scenes for training and 647 scenes for development data.

4.2. Evaluation Metrics

We used four metrics for evaluating our model, baseline MBSTOI, and HASPI: Pearson correlations (ρ), root-mean-square error (RMSE), F1 score (F1), and area under the curve (AUC) [22]. The speech intelligibility prediction model generates an intelligibility score ranging from 0 to 100, as defined in the CPC1 challenge [4]. Then, we converted the scale of baseline MBSTOI and HASPI from 0–1 to 0–100 by performing the RMSE minimization using a sigmoid function. We calculated the ρ and RMSE of the predicted scores with the actual correctness of the subjective listening test. The F1 and AUC scores were obtained using binary classification (high and low). The score is classified as high when it is larger than 50 (middle point of 0–100); otherwise, it is classified as being a low score.

Table 1: Evaluation results of several speech intelligibility prediction models: MBSTOI (Baseline), HASPI left ear (HASPI (left)), HASPI right ear (HASPI (right)), and our proposed model with HIC predictor (OBISHI+HIC) and without HIC predictor (OBISHI).

Dataset	Method	Track 1 (close-set)				Track 2 (open-set)			
		ρ	RMSE	F1 (%)	AUC (%)	ρ	RMSE	F1 (%)	AUC (%)
Dev	**Baseline**	0.63	33.65 ± 1.42	81.01	76.11	0.48	33.77 ± 0.92	84.57	67.18
	HASPI (left)	0.67	36.07 ± 1.34	73.13	71.91	0.43	43.58 ± 1.02	52.91	58.15
	HASPI (right)	0.67	35.57 ± 1.34	73.10	72.27	0.45	42.40 ± 1.01	57.16	58.67
	OBISHI	0.70	25.97 ± 1.21	**88.55**	85.23	**0.60**	**22.81 ± 0.84**	**90.92**	**77.19**
	OBISHI+HIC	**0.77**	**23.97 ± 1.16**	88.21	**86.13**				
Test	**Baseline**	0.62	28.52 ± 0.58	81.83	75.74	0.53	36.52 ± 1.35	68.39	68.74
	HASPI (left)	0.60	37.72 ± 0.60	68.33	68.56	0.57	37.87 ± 1.20	67.88	68.58
	HASPI (right)	0.60	37.66 ± 0.60	68.33	68.56	0.55	38.61 ± 1.23	67.05	67.99
	OBISHI	**0.68**	**27.86 ± 0.54**	85.04	80.72	**0.67**	**28.29 ± 1.06**	**82.90**	**78.69**
	OBISHI+HIC	0.41	37.19 ± 0.72	**85.16**	**87.11**				

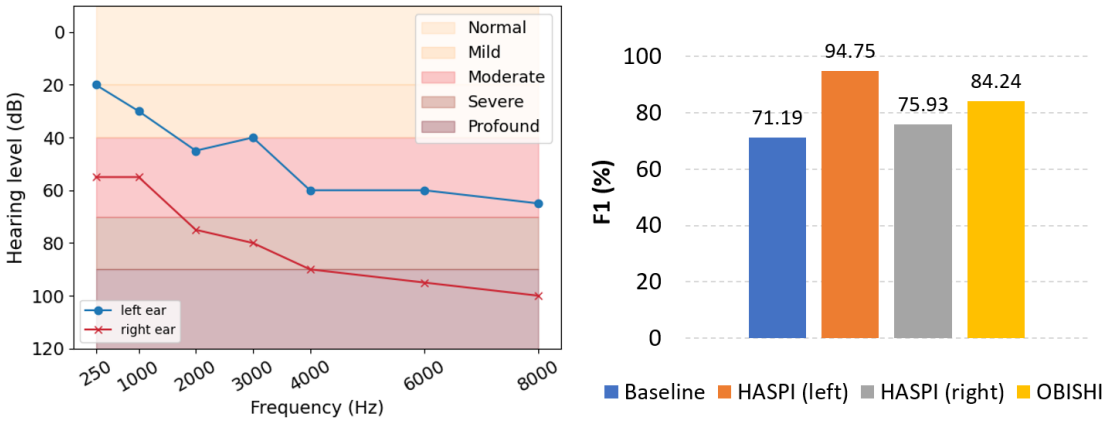


Figure 4: A case study on listener L0217. The left panel shows the audiogram of listener L0217. The right panel shows the speech intelligibility predictions results using comparative methods in terms of F1 score.

4.3. Results

Table 1 shows the total evaluation results of three models in both the development and testing phases. In general, our method improved the intelligibility prediction compared to the baseline and HASPI models.

[Development phase] In comparison with the baseline model, our prediction method significantly reduces the RMSE by approximately 10% and classification accuracy (F1 and AUC) by more than 5% for both tracks. An additional hearing characteristic predictor could also slightly improve the close-set scenario prediction. The performance of HASPI is generally not as high as that of the baseline and proposed models.

[Testing phase] Overall results during the testing phase also indicate that our proposed method has the best performance. However, although the classification accuracy could be improved, the additional HIC predictor in the proposed model (OBISHI+HIC) increased the RMSE and reduced the correlation. We predicted that this issue was caused by the rising number of missing hearing characteristics data occurring in the test set of track 1, which is larger than the development set. Although the imputation approach has been applied to the missing data, the model may fail to predict the relevant hearing characteristics beneficial for predicting speech intelligibility scores. Without the HIC predictor, our proposed method can improve the prediction accuracy, especially in track 2.

[A case study] We also plotted the classification prediction results and the audiogram of a specific HI listener 'L0217' in Fig. 4. We chose this listener because the hearing condition of the left ear is different than the right ear. The audiogram in Fig. 4 revealed listener L0217 has profound hearing loss in the right ear but a moderate hearing loss in the left ear at a higher frequency ($>$ 4 kHz). This condition is well represented by the HASPI model, where the left ear is better than the right ear. Meanwhile, the proposed model can balance the intelligibility prediction of both ears (F1 = 84.24%) better than the baseline model (F1 = 71.19%).

5. Conclusions

This paper proposed an objective binaural intelligibility score for the hearing impaired, OBISHI. The OBISHI belongs to an intrusive metric that considers the HI characteristics for predicting the speech intelligibility score. Additionally, we utilized an ASR system to infer the difficulty of the utterances in an NH condition. We integrated the MSBG hearing loss model with our constructed GTFB-based predictor model in the intelligibility model. The evaluation was conducted using a training test split method on two tracks (close-set and open-set). We also compared the predicted intelligibility score of the baseline MBSTOI and HASPI with the actual correctness from the listening test. The results showed that our method could significantly improve the prediction of the baseline MBSTOI and HASPI for both close-set and open-set tracks. In addition, our proposed method significantly improved the speech intelligibility prediction when the listener has different hearing impaired conditions of left and right ears compared to the baseline method.

6. Acknowledgements

This work was supported by the SCOPE Program of Ministry of Internal Affairs and Communications (no. 201605002), a Grant-in-Aid for Scientific Research (B) (no. 21H03463), and a JSPS KAKENHI grant (no. 22K21304). This work was also partially supported by the Japan Society for the Promotion of Science (JSPS) KAKENHI (grant numbers 22H04860 and 22H00536) and JST AIP Trilateral AI Research, Japan (grant number JPMJCR20G6).

7. References

[1] C. J. Plack, V. Drga, and E. A. Lopez-Poveda, "Inferred basilar-membrane response functions for listeners with mild to moderate sensorineural hearing loss." *Journal of the Acoustical Society of America*, vol. 115 4, pp. 1684–95, 2004.

[2] B. C. J. Moore, D. A. Vickers, C. J. Plack, and A. J. Oxenham, "Inter-relationship between different psychoacoustic measures assumed to be related to the cochlear active mechanism." *Journal of the Acoustical Society of America*, vol. 106 5, pp. 2761–78, 1999.

[3] M. Munro and T. Derwing, "Processing time, accent, and comprehensibility in the perception of native and foreign-accented speech," *Language and speech*, vol. 38 (3), pp. 289–306, 07 1995.

[4] J. Barker, M. Akeroyd, T. J. Cox, J. F. Culling, J. Firth, S. Graetzer, H. Griffiths, L. Harris, G. Naylor, Z. Podwinska, E. Porter, and R. V. Munoz, "The 1st Clarity Prediction Challenge: A machine learning challenge for hearing aid intelligibility prediction," in *INTERSPEECH 2022*. ISCA, 2022.

[5] J. Kates and K. Arehart, "The hearing-aid speech perception index (HASPI)," *Speech Communication*, vol. 65, 11 2014.

[6] J. M. Kates and K. H. Arehart, "The hearing-aid speech perception index (HASPI) version 2," *Speech Communication*, vol. 131, pp. 35–46, 2021.

[7] A. H. Andersen, J. M. de Haan, Z. Tan, and J. H. Jensen, "Refinement and validation of the binaural short time objective intelligibility measure for spatially diverse conditions," *Speech Commun.*, vol. 102, pp. 1–13, 2018.

[8] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time–frequency weighted noisy speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2125–2136, 2011.

[9] A. H. Andersen, J. M. de Haan, Z.-H. Tan, and J. Jensen, "A method for predicting the intelligibility of noisy and non-linearly enhanced binaural speech," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 4995–4999.

[10] Y. Nejime and B. Moore, "Simulation of the effect of threshold elevation and loudness recruitment combined with reduced frequency selectivity on the intelligibility of speech in noise," *Journal of the Acoustical Society of America*, vol. 102, pp. 603–15, 08 1997.

[11] B. C. J. Moore and B. R. Glasberg, "Suggested formulae for calculating auditory-filter bandwidths and excitation patterns." *Journal of the Acoustical Society of America*, vol. 74 3, pp. 750–3, 1983.

[12] J. Kiessling, "Current approach to hearing aid evaluation," *Canadian Journal of Speech-Language Pathology and Audiology*, vol. 17, no. 4, pp. 39–49, 1993.

[13] D. M. Harris and P. Dallos, "Forward masking of auditory nerve fiber responses," *Journal of neurophysiology*, vol. 42, no. 4, pp. 1083–1107, 1979.

[14] S. Graetzer, J. Barker, T. J. Cox, M. Akeroyd, J. F. Culling, G. Naylor, E. Porter, and R. V. Muñoz, "Clarity-2021 challenges: Machine learning challenges for advancing hearing aid processing," in *Interspeech 2021*. ISCA, 2021, pp. 686–690.

[15] K. Andersson, L. Andersen, J. Christensen, and T. Neher, "Assessing Real-Life Benefit From Hearing-Aid Noise Management: SSQ12 Questionnaire Versus Ecological Momentary Assessment With Acoustic Data-Logging," *American Journal of Audiology*, vol. 30, 12 2020.

[16] W. Whitmer, P. Howell, and M. Akeroyd, "Proposed norms for the Glasgow hearing-aid benefit profile (GHABP) questionnaire," *International journal of audiology*, vol. 53, 02 2014.

[17] E. V. den Borre, S. Denys, A. van Wieringen, and J. Wouters, "The digit triplet test: a scoping review," *International Journal of Audiology*, vol. 60, no. 12, pp. 946–963, 2021.

[18] N. Tomashenko, B. M. L. Srivastava, X. Wang, E. Vincent, A. Nautsch, J. Yamagishi, N. Evans, J. Patino, J.-F. Bonastre, P.-G. Noé, and M. Todisco, "The VoicePrivacy 2020 Challenge evaluation plan," 2020.

[19] V. Peddinti, D. Povey, and S. Khudanpur, "A time delay neural network architecture for efficient modeling of long temporal contexts," in *INTERSPEECH 2015*, 2015, pp. 3214–3218.

[20] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *2015 IEEE ICASSP*, 2015, pp. 5206–5210.

[21] M. Unoki and M. Akagi, "A method of signal extraction from noisy signal based on auditory scene analysis," *Speech Communication*, vol. 27, pp. 261–279, 4 1999.

[22] D. Freedman, R. Pisani, and R. Purves, "Statistics (international student edition)," *Pisani, R. Purves, 4th edn. WW Norton & Company, New York*, 2007.

A Human Annotation Guide for Mental Health Speech Collections

Brian Stasak[1,2], Julien Epps[2], Mark Larsen[1], and Helen Christensen[1]

[1]Black Dog Institute, UNSW, Sydney, NSW – Australia
[2]School of Elec. Eng. & Telecomm., UNSW, Sydney, NSW – Australia
b.stasak@unsw.edu.au, j.epps@unsw.edu.au, mark.larsen@blackdog.org.au, h.christensen@blackdog.org.au

Abstract

While large amounts of recorded speech audio data are collected for medical analysis, there is not a compact guideline available that outlines which human-rated annotations are important to consider when analyzing speech from individuals with potential mental illness. Herein, an annotation guideline is proposed that highlights fifty-two different speech-audio recording transcription considerations, including several new ones related to voice accent, prosodic, intelligibility, quality, auxiliary behavior, task compliance, disfluency, and noise factors. Further, a free, new software tool for recording recommended speech-based mental health annotations is provided to help scientists who may be unfamiliar with speech-audio data collection.

Index Terms: digital medicine, survey, transcription, voice

1. Introduction

Human-rated annotations derived from audio speech recordings are vital in understanding communication behavior, protocol design, real-world compliance, and noise factors that may adversely impact the quality of recorded audio. Speech annotations provide specific details about abnormal speech-language characteristics exhibited by individuals experiencing mental ill-health and may be used to help screen or predict mental health outcomes. While automatic annotation methods (e.g., automatic speech recognition, signal-to-noise ratio, syllable durations) are available to evaluate speech recordings, the human-rated annotation approach is still the 'gold standard' [1]. Moreover, automatic annotation approaches require kernels of data with human-rated annotations to generate robust modeling and to establish system test performance baselines.

While many medical speech datasets concerning neurological disorders (e.g., aphasia, dementia, Parkinson's disease) frequently contain transcripts including linguistic-level part-of-speech annotations, the few scarcely available mental health speech datasets omit this level of information [2]. Moreover, mental health speech datasets rarely include subjective human-rated insights, such as voice quality, paralinguistic traits, auxiliary behaviors, task compliance, and noise annotations. It is believed that these types of annotations may further reveal abnormalities in emotion and language patterns in individuals with mental health disorders [2].

While efforts have been made to create uniform, widely accepted annotation guidelines, most mental health research speech corpora still contain different sets of annotation taxonomies and procedures [3]. Many data collections provide vague details on the human annotation process, making it more difficult to replicate the annotation standards in future studies [1-3]. Thus, there is a need for a uniformly coded metadata annotation system and a convenient easy-to-use human-labeling software tool for mental health speech data collections. Recently, voice quality, prosodic, and speech disfluency annotations have proven useful in detecting individuals with depression [4-9]. Other reported speech annotations, such as auxiliary vocal behaviors, task compliance, and noise have been explored in the mental health literature [10-12]. Speech-language and audio-signal related annotations are of substantial interest since modern mental health smartphone collection speech recordings are often conducted in a real-world natural home environment without a clinician/administrator present.

In this paper, annotations based on six different speech-audio related categories were chosen based on previous speech-based mental health and voice studies [4-12]. Per speech annotation category, this paper includes a brief literature review, a description of known connections to mental illness, and proposed annotation approaches to help capture relevant speech behaviors for further analysis. A newly proposed human annotation system and uniform software tool provides groundwork for scientists labeling speech-audio metadata to help discover new speech-related sub-symptoms in mental illness. Annotated dataset results may lead to the development of automated near-real-time annotation systems, which can be implemented on a large scale – reducing annotation time and costs. Further, automated annotation systems could also help to reduce possible human-rater bias, improve health annotator safety, and allow greater patient anonymity.

2. Speech Annotation Categories

2.1. Voice Quality

There are many factors that can impact voice quality, such as an individual's height, stress, personal habits (e.g., alcohol consumption, smoking, vaping), amount of vocal use/abuse, trauma injury, and diagnosed illness (e.g., voice, neurological, psychogenic disorders). In general populations, it is estimated that up to 9% of the population has a clinical voice disorder [13, 14]. Recent speech-based studies [8, 9] have discovered strong correlations between mental health disorders and voice quality. For example, [8] found that when given read sentence tasks, individuals with recent suicidal ideation or attempt exhibited significantly poorer voice quality scores than non-suicidal individuals. Further, in [9], it was found that despite different speaker age groups (e.g., 18-34, 35-48, 49-62, 63-79), individuals with higher depression severity scores demonstrated poorer voice quality.

Individuals' perception of their voice quality often reflects their degree of self-esteem and recognized changes in voice due to medical conditions [15]. Based on human perceptual ratings, undesirable vocal characteristics include unusual nasality and monotony [16]. For example, according to [17], individuals

with chronic sinusitis and nasal polyp disorders have nearly twice the risk of clinical depression disorders than healthy controls possibly in part due to self-esteem, voice quality, and other contributing factors (e.g., sleep habit, stress).

Studies [18, 19] have highlighted that the GRBASI subjective voice quality assessment is widely used in the field of clinical voice evaluation and research. Currently, the GRBASI is the evaluation standard, even when compared to more objective automatic approaches [18]. Further, GRBASI ratings were shown to have good interrater reliability across many different types of voice disorders [19].

While all individuals have unique resonance/production factors that result from a blend of voice quality attributes, most are normal/slight and do not negatively impact verbal expressiveness and intelligibility. The GRBASI utilizes a four-point scale (e.g., 0-normal, 1-slight 2-moderate, 3-severe) and it requires a pre-trained listener to subjectively rate a speaker based on six different voice quality attributes. Individual GRBASI scores are then added together to produce a GRBASI composite score. The higher the GRBASI composite score, the poorer an individual's voice quality. The six different GRBASI voice quality attributes are described briefly as follows: (1) *grade*: hoarseness, raspy; (2) *roughness*: low frequency vibration irregularity; (3) *breathiness*: air escaping, leakage, whispery; (4) *asthenia*: loss of power, weakness; (5) *strain*: hyper-functional phonation, rattle; and (6) *instability*: inconsistent quality, voice fluctuation. For more details regarding the GRBASI voice quality assessment, see [18, 19].

In addition to the GRBASI voice quality ratings, another important voice quality is nasal resonance. Nasality is unrelated to laryngeal function (e.g., vocal folds) and requires independent activation of the nasopharynx. In English, only the /n, m, ŋ/ sounds have nasality. Both hypo/hyper nasality can cause a loss of speech clarity and linguistic stress between syllables [20]. Hyponasality means that for nasal sounds /n, m, ŋ/, no acoustic vocal energy passes through the nasal passage (i.e., no air can escape via the nostrils). This creates a voice quality found in individuals with a cold or allergies with a stuffed-up nose. Hypernasality means that vocal energy is accidently escaping into the nasal passage while making all sounds, rather than only nasal sounds /n, m, ŋ/. In hypernasality, there is a faulty control in palato-pharyngeal valving, which results in excessive nasalization of vowels and non-nasal consonants. Hypernasality causes a voice quality often found in individuals with Down Syndrome (i.e., due to alteration in velopharyngeal sphincter mobility and epipharynx narrowing) and cleft palate.

In addition to the recommended GRBASI voice quality assessment, a nasality annotation is also proposed using three possible presence indicators (0-normal, 1-hyponasality, 2-hypernasality). A score of '0' is normal nasopharynx function without noticeable hypo/hyper traits.

2.2. Accent, Prosody, Rate, and Confidence

Refugee/immigrant mental health studies have shown a high risk of mental distress due to various traumas (e.g., war, natural disaster, famine) and struggles (e.g., crime, discrimination, financial, employment) when compared with the general local population [21]. It has also been reported that some ethnicities/cultures have historically and significantly higher incidence of mental health issues. For example, in the United States, Native Americans have the highest rates of any mental health diagnosis than any other ethnic population [22]. Due to the high prevalence of mental health concerns in refugee/immigrant populations and potential non-native language skill deficits, speech accent is an important factor to consider when analyzing speech recordings.

Early subjective observational paralinguistic depression studies [23, 24] by clinicians described patients with depression as having abnormal speech patterns, such as weaker loudness, slower rate, flattened pitch, uniform rhythm, less verbosity, and an unusual lifeless or hollow sounding timbre. Modern acoustic speech studies [7, 25] have also indicated decreases in depressed speakers' loudness, stress-prosodic characteristics, verbal fluency, and overall rate-of-speech. A recent study [7] found that depressed populations exhibit more uniform syllable durations and flatter amplitude dynamics than healthy populations, resulting in less prosodic word stress and overall poorer intelligibility of speech. Under experimental conditions, researchers found that that individuals with social anxiety disorder subsequently exhibited decreased vocal confidence in contrast to a control group [26].

Rate-of-speech annotations have been explored in mental health studies [27-29], which have indicated changes in speech fluency during depressive, anxiety, or schizophrenic episodes. It is known that depression can impair fine-motor skills and lead to symptomatic psychomotor retardation (i.e., decreased rate-of-speech, increased syllable lengths, shorter sentence utterances) or agitation (i.e., excessive rapid gesturing, accelerated motor activity, and verbose activity).

It is expected in a speech collection that most speakers will use a native accent. Thus, a native accent binary annotation is proposed using the indicators 0-native and 1-foreign. The native accent annotation rating is subjectively based on the native production of a speaker's phonemes (i.e., often non-native speakers will use their native language phoneme prototypes). A new degree of accentedness annotation rating range is also proposed by the following: 0-native; 1-light foreign accent; 2-medium foreign accent; and 3-heavy foreign accent. Further, a new speech intelligibility rating based on five ratings is also proposed using: 0-very low; 1-low; 2-moderate; 3-high; and 4-very high. A 4-very high speech intelligibility rating is described as clear flawless enunciated speech (i.e., high comprehension), whereas a 0-very low intelligibility rating implies very little of what was spoken could be comprehended by a listener (i.e., accurately repeated/written).

The concept of natural speech continuum is rated using a binary whereby any recording with a natural stream of speech (i.e., without unnatural pauses/breaths) is given a '0'. Conversely, a recording with a speech pattern that has abnormal disruptions is given a '1' (i.e., abnormally broken or disrupted). Per recording, similarly to [8], an overall rate-of-speech is recommended using the following ratings: 0-very slow; 1-slow; 2-moderate; 3-fast; and 4-very fast. The rate-of-speech is rated based on the average of the entire recording. If the participant speaks very slowly during one sentence and then very fast for the next, a rating of 'moderate' (2) is given for the recording.

A prosodic annotation rating indicates the level of paralinguistic dynamics per recording. Prosodic annotation ratings are proposed as follows: 0-very flat; 1-flat; 2-moderate; 3-dynamic; and 4-very dynamic. Similar to the rate-of-speech annotation, the prosody annotation is rated based on the average of the entire recording. An example of 0-very flat prosody annotation rating includes a speaker that uses a monotone voice with minimal amplitude variation, rate of speech, and pitch contour (e.g., Ben Stein). On the contrary, a 4-very dynamic prosody annotation rating includes a speaker that uses excessively wide amplitude, rate of speech, and pitch contour ranges (e.g., Robin Williams).

The speech task confidence rating is related to how confident a speaker sounds (i.e., did the participant sound more assertive or unsure during his/her speech task recording?). A speaker's confidence is inferred by a combination of vocal amplitude, speed, directness, dominance, and recorded response fluidity [31]. The proposed task confidence annotation relies on five ratings (0-very low; 1-low; 2-moderate; 3-high; 4-very high). An example of a 0-very low rating is an individual that perceptually sounds uncertain of during an utterance. A 4-very high rated confident individual will produce utterances that sound more like a factual statement with increased speed/loudness. Note that an individuals may sound confident even though they may utter incorrect responses during speech tasks (e.g., picture naming, read sentences, Stroop color test).

2.3. Speech Disfluencies

Individuals with a serious mental illness (e.g., bipolar/unipolar depression, schizophrenia) struggle in nearly every aspect of speech production when compared with control participants [14]. Studies [4-8] have shown that depression can be detected through individuals' speech disfluencies. For example, analysis of read speech tasks demonstrated statistically significant feature differences in speech disfluencies, whereby when compared with non-depressed speakers, depressed speakers showed relatively higher recorded frequencies of hesitations (55% increase) and speech errors (71% increase) [7]. Another investigation of speech recordings taken from inpatients with suicidal ideation and suicide attempt had approximately twice as many hesitations and four times as many speech errors when compared with individuals in a control group [8]. Also, when compared with the control groups, it was shown in both studies [7, 8] that individuals with depression tended to incorrectly substitute words without self-corrections.

Disfluency annotations have been applied in previous speech-based mental health studies [7, 8]. Speech repeats occur at a syllable, word, or phrase level. Raw counts are annotated to record the exact number of repeats per type. In some instances, it is possible for a spoken utterance to include a count for all three repeat types. Hesitations include two proposed annotated forms: non-speech (e.g., pause) and speech (e.g., vocal held sound). A non-speech pause hesitation typically has an abnormal gap of silence (i.e., not due to end of phrase, breath, or emphasis) that occurs abruptly, disrupting the flow of what is being said. A vocalized speech hesitation is one which is achieved by abnormally holding a sound for an unusually longer duration during an utterance. A vocal speech hesitation will frequently occur when an individual is unsure or cues that there is more to say (i.e., holding speaker-turn-taking dominance, time for recollecting thoughts).

The speech error annotation is a total raw count summary of any instance of a speech error, including *all* disfluency types (e.g., grammar, phonological, repetitions, hesitations, substitutions, deletions, insertions, malapropisms). Individuals with non-native accents or English-as-second language should not have phonological disfluency errors tracked due to normal phoneme-mapping shifts (e.g., Grimm's Law). A raw count of specific speech error types is further provided for substitutions (e.g., '*took*' / '*take*'), deletions (e.g., '*cars*' / '*car_*'), insertions (e.g., '*He left yesterday*' / '*He also left yesterday*'), and malapropisms (e.g., '*amuse*' / '*mouse*'). Substitutions are usually a variant of the target word, target sounds, and are the same word class (e.g., noun, verb, adjective), whereas malapropisms involve the substitution of the entire word often with a nonsensical effect. Another annotation proposed is self-corrections [7, 8], wherein for every self-correction opportunity evaded a raw count is recorded. It is easiest to score self-correction during read speech tasks because it is known exactly what the speaker should have said.

2.4. Auxiliary Vocal Behaviors

Auxiliary vocal behaviors are common during speech production (e.g., coughs, laughs, throat clearing, sighs, yawns). Individuals with mental health disorders often have higher incidence of additional illness (e.g., comorbidity) [31], which may impact speech production and potentially increase any number of auxiliary vocal behaviors. Abnormal sleep patterns are a common symptom among individuals with depression, whereby roughly 75% have insomnia symptoms [32].

Self-comments (i.e., externalized monologue) are generally a coping strategy used during high cognitive load competition-type tasks. However, under certain conditions, excessive self-comment verbalizations can be an indicator of mental health disorders (e.g., schizophrenia, hallucinatory episode) [34]. For example, during a Stroop color test, some individuals make comments aloud about their test performance (e.g., "*Green*, *blue*, *yellow*, *white*, black, *oh I really messed up there*, *yellow*, *damn it*."). Formulaic language is common in everyday discourse, where it contributes to nearly a quarter of all conversational speech [10, 12]. Formulaic language includes conventional word expressions, proverbs, idioms, expletives, hedges, bundles, and fillers.

Similar to [12], reporting a raw count for common word filler types (e.g., *ah*, *er*, *mm*, *uh*, *um*, *so*, *like*, *you know*) additional proposed raw count annotations are self-comments, expletives (e.g., swears), and speech auxiliary behavior types (e.g., coughs, laughs, throat clearing, sighs, yawns).

2.5. Task Compliance

Task compliance is rarely reported in speech-based data collections. Open-source speech datasets that are frequently relied upon for experimental analysis and publications often contain recordings wherein subjects did not follow the given directions properly. Task compliance is useful in understanding what percentage of recordings were completed correctly, abnormal verbal responses, and also whether or not some speaker subsets have higher compliance than others. The task compliance can also provide better insight in terms of how well a task is explained for participants, and further, whether the task instructions should be revised to increase task compliance.

The newly proposed task compliance annotation rating indicates whether the task directions were properly 0-completed or 1-non-compliant. Some examples of a marking of 1-non-compliance include: a speaker who says nothing during the recording; a speaker that says something else than what the task specifies; the speaker has another individual speak on his/her behalf (i.e., can determine based on gender/age factors); and the speaker does not fully complete the given task. Some speech tasks (e.g., held vowel, diadochokinetic, word fluency, Stroop color test, read sentence) require the speaker to generate specific target examples of related word tokens. Therefore, a proposed task-specific count annotation is also recommended (i.e., raw target word count versus total word count).

2.6. Noise Factors

During recorded speech samples, especially those collected outside of a laboratory setting (e.g., natural environment), background and channel noise level are important to annotate.

Studies [34, 35] have demonstrated that individuals living in home environments with noise pollution have greater incidence of poorer health and quality of life. Noise annotations can reveal information that may be related to higher incidence of mental health disorders (e.g., noise pollution, ability to focus on single task). This noise-based metadata can also be useful to build noise-specific modeling (i.e., increase noise robustness, feature reliability), test specific automated system performance given different noise types/levels, and help explain unusual statistical feature analysis results found in mental health populations.

For background noise level, four ratings are proposed (0-none; 1-minimal; 2-moderate; 3-severe). A background noise level of 0-none is a very quiet environment without background noise, wherein the signal-to-noise ratio is very strong, and the speech is clear. A background noise level of 1-minimal is where the signal-to-noise ratio is still strong, but there may be a small amount of background noise, and it does not impede intelligibility. A background noise level of 2-moderate is when the noise level has increased to nearing the level of the speech amplitude, possibly making some part/s of speech harder to comprehend. A background noise level of 3-severe is when the background noise is so loud it surpasses the speech signal level, making the speech difficult to understand.

A proposed noise duration annotation is as follows: 0-none; 1-intermittant; 2-continous; and 3-mixed (i.e., both intermittent and continuous). Further, specific noise type annotations are suggested the following: 0-none; 1-Radio/TV; 2-babble; 3-machinery; and 4-other. Further, the background speaker annotation (i.e., secondary speaker/s) is a binary value of either '0-absent' or '1-present'. Device noise annotation related to a recording device channel issues (e.g., bit-loss, buzz, clipping, device error, hum) is also proposed using the following scale: 0-none; 1-minimal; 2-moderate; 3-severe; and 4-very severe.

Similar to the standard mean opinion score (MOS) annotation based on audio listening speech-to-noise signal quality [36], a new rating scale for rating speech-to-noise signal quality is proposed: 1-bad (very annoying and objectionable); 2-poor (annoying but not objectionable); 3-fair (perceptible and slightly annoying); 4-good (just perceptible, but not annoying; and 5-excellent (level of distortion imperceptible). The speech-to-noise signal quality rating deals with how perceptually pleasing a recording is in terms of signal-to-noise ratio, whereas the previously mentioned noise level annotation is a perception of the clarity of the speech signal despite possible background noise or compression loss. A secondary activity annotation is also proposed as 0-none, 1-eating/drinking; 2-driving/in-vehicle; and 3-other.

3. Annotation Software Design

A new survey tool for human annotations and audio listening capability was created using python code which is freely available at: https://www.researchgate.net/profile/brian_stasak. This includes a GUI for batch processing of audio file recording playback and file-by-file survey reporting. A digital survey was created that included all 52 annotations described in this paper. Further, metadata regarding the annotator, such as gender, age, first-language, headphone/speaker type, years annotator experience, and listening environment loudness, were also included in the digital survey app. The app exports the annotation survey information into a .CVS file format for later analysis. Currently, this app tool is helping to annotate an ongoing Black Dog Institute mental health data collection that to-date includes over 7k recordings from more than 1k school-aged participants in naturalistic conditions [37].

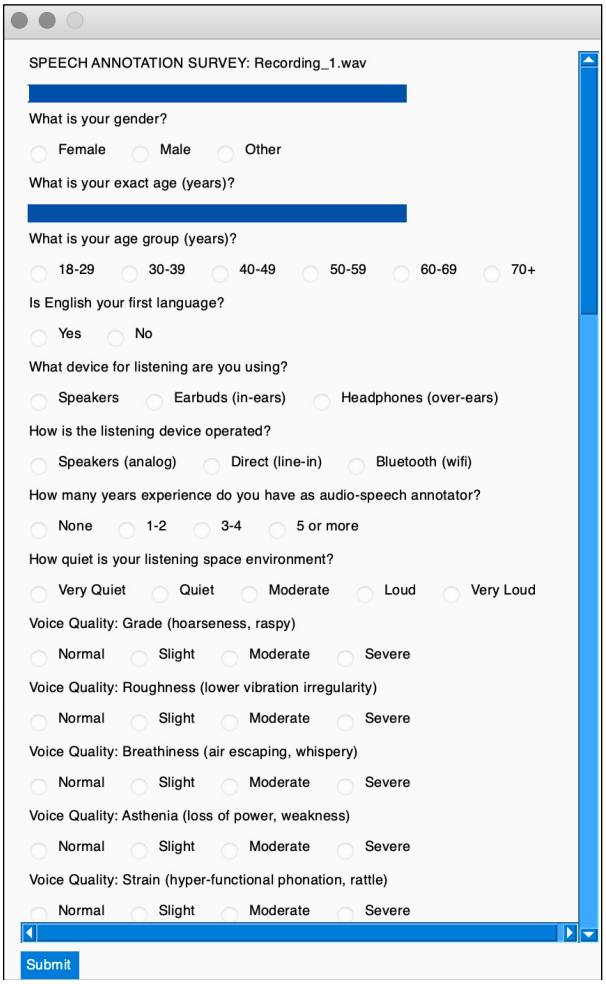


Figure 2: Image of python-based human-rated annotation GUI for speech-audio playback listening and survey notations.

4. Conclusion

While previous speech-based mental health data collections have evaluated one or two human annotation categories, none have conducted a comprehensive examination that includes *all* six categories described in this paper. Due to the limited number of speech-based mental health databases available, it is imperative that maximum information is gleaned for analysis. This proposed speech-based mental health annotation guide was streamlined into a novel annotation survey tool, which is useful in unifying annotation standards and comparing future datasets. While transcription annotation software (e.g., Praat, SpeechTools) is commonly used in audio-speech notation, annotation software guides readily available that implement broad categorical annotations are limited. This annotation tool is helpful to non-audio-speech experts who are interested in evaluating recorded audio (e.g., medical professionals, linguists); and/or digital health experts designing automatic speech-based machine learning models, whereby human-rated annotations are required to generate 'ground-truth' performance baseline and test model parameter optimization.

5. Acknowledgements

Funding for this project came from a NHMRC Project Grant Awarded to HC (APP1120646).

6. References

[1] Li, X., Liu, H., Kury, F., Yuan, C., Butler, A., Sun, Y., Ostropolets, A., Xu, H., and Weng, C., "A comparison between human and NLP-based annotation of clinical trial eligibility criteria text using the OMOP common data model", In: Proc. AMIA Joint Summits on Transl. Science, pp. 394–403, 2021.

[2] Li, Y., Ding, H., and Li, D., "Speech databases for mental disorders: a systematic review", General Psych., vol. 32(3), pp. 1–10, 2019.

[3] Delais-Roussarie, E. and Post, B., "Corpus annotation: methodology and transcript systems", The Oxford Handbook of Corpus Phonology, Oxford University Press, Oxford - UK, 2014.

[4] Esposito, A., Esposito, A.M., Likforman-Sulem, L., Maldonato, M.N., Vinciarelli, A., "On the significance of speech pauses in depressive disorders: results on read and spontaneous narratives", Recent Advances in Nonlinear Speech Processing, Smart Innovation, Systems and Technologies, vol. 48, Springer, 2016.

[5] Oxman, T. E., Rosenberg, S. D., Schnurr, P. P., and Tucker, G. J., "Diagnostic classification through content analysis of patients' speech", American J. of Psych., vol. 145(4), pp. 464–468, 1988.

[6] Rubino, A., D'Agostino, L., Sarchiola, L., Romeo, D., Siracusano, A., and Docherty, N.M., "Referential failures and affective reactivity of language in schizophrenia and unipolar depression", Schizophrenia Bulletin, vol. 37(3), pp. 554–560, 2011.

[7] Stasak, B., Epps, J. and Goecke, R., "Automatic depression classification based on affective read sentences: opportunities for text-dependent analysis", Speech Comm., vol. 115, pp. 1–14, 2019.

[8] Stasak, B., Epps, J., Schatten, H.T., Miller, I.W., Provost, E.M. and Armey, M.F., "Read speech voice quality and disfluency in individuals with recent suicidal ideation or suicide attempt", Speech Comm., vol. 132, pp.10–20, 2021.

[9] Stasak, B., Joachim, D. and Epps, J., "Breaking age barriers with automatic voice-based depression detection, IEEE Pervasive Computing, pp. 1–5, 2022.

[10] Bridges, K.A., "Prosody and formulaic language in treatment-resistant depression: effects of deep brain stimulation", Doctoral Thesis - New York University, 2014.

[11] Pope, B., Blass, T., Siegman, A.W., and Raher, J., "Anxiety and depression in speech", J. of Consulting and Clinical Psych., vol. 35(1), pp. 128–138, 1970.

[12] Stasak, B., Epps, J., and Cummins, N., "Depression prediction via acoustic analysis of formulaic word fillers", Polar, vol. 77(74), pp. 230–234, 2016.

[13] Roy, N., Merrill, R.M., Gray, S.D., and Smith, E.M., "Voice disorders in the general population: prevalence, risk factors, and occupational impact", The Laryngoscope, vol. 115(11), pp. 1988–1995, 2005.

[14] Cohen, A.S., McGovern, J.E., Dinzeo, T.J., and Covington, M.A., "Speech deficits in serious mental illness: a cognitive resource issue?", Schizophrenia Res., vol. 160(1-3), pp.173–179, 2014.

[15] Berto, V., "The relationship between perceptions of vocal quality and function with self-esteem in older adults", Theses and Dissertations - Illinois State University, pp. 1–65, 2018.

[16] Zuckerman, M. and Miyake, K., "The attractive voice: what makes it so?", J. Nonverbal Behavior, 17(2), pp. 119–135, 1993.

[17] Kim, J.Y., Ko, I., Kim, M.S., Yu, M.S., Cho, B.J., and Kim, D.K., "Association of chronic rhinosinusitis with depression and anxiety in a nationwide insurance population", JAMA Otolaryngology–Head & Neck Surgery, vol. 145(4), pp. 313–319, 2019.

[18] Barsties, B. and De Bodt, M., "Assessment of voice quality: current state-of-the-art", Auris Nasus Larynx, vol. 42(3), pp. 183–188, 2015.

[19] Yamaguchi, H., Shrivastav, R., Andrews, M.L., and Niimi, S., "A comparison of voice quality ratings made by Japanese and American listeners using the GRBAS scale", Folia Phoniatrica et Logopaedica, vol. 55(3), pp. 147–157, 2003.

[20] Oren, L., Kummer, A., and Boyce, S., "Understanding nasal emission during speech production: a review of types, terminology, and causality", Cleft Palate Craniofac. J., vol. 57(1), pp. 123–126, 2020.

[21] Pumariega, A.J., Rothe, E., and Pumariega, J.B., "Mental health of immigrants and refugees, Comm. Mental Health J., vol. 41(5), pp. 581–597, 2005.

[22] Coleman, K.J., et al., "Racial/ethnic differences in diagnosis and treatment of mental health conditions across healthcare system participants in the mental health research network", vol. 67(7), pp. 749–757, 2016

[23] Newman, S. and Mather, V.G., "Analysis of spoken language of patients with affective disorders", American J. of Psych., vol. 94(4), pp. 913–942, 1938.

[24] Stinchfield, S.M., "Speech disorders: a psychoanalytical study of the various defects in speech", New York, NY - USA, 1933.

[25] Alpert, M., Pouget, E.R., and Silva, R.R., "Reflections of depression in acoustic measures of the patient's speech", J. Affective Disorders, vol. 66(1), pp. 59–69, 2001.

[26] Gilboa-Schechtman, E., Galili, L., Sahar, Y., and Amir, O., "Being "in" or "out" of the game: subjective and acoustic reactions to exclusion and popularity in social anxiety", Frontiers in Human Neuroscience, vol. 8, pp. 147–160, 2014.

[27] Flint, A.J., Black, S.E., Campbell-Taylor, I., Gailey, G.F., and Levinton, C., "Abnormal speech articulation, psychomotor retardation, and subcortical dysfunction in major depression", J. Psychiatric Res., vol. 27(3), pp. 309–319, 1993.

[28] Kuny, S.T. and Stassen, H.H., "Speaking behavior and voice sound characteristics in depressive patients during recovery", J. of Psychiatric Res., vol. 27(3), pp. 289–307, 1993.

[29] Nilsonne, A., "Speech characteristics as indicators of depressive illness", Acta Psychiatrica Scandinavica, vol. 77, pp. 253–263, 1988.

[30] Kimble, C.E. and Seidel, S.D., "Vocal signs of confidence", J. Nonverbal Behavior, vol. 15(2), pp. 99–105, 1991.

[31] Goodell, S., Druss, B.G., Walker, E.R., and Mat, M.J.R.W.J.F.P., "Mental disorders and medical comorbidity", report, Robert Wood Johnson Foundation, vol. 2(21), pp. 1–6, 2011.

[32] Nutt, D., Wilson, S., and Paterson, L., "Sleep disorders as core symptoms of depression", Dialogues Clinical Neurosci., vol 10(3), pp. 329–336, 2008.

[33] Tovar, A., Fuentes-Claramonte, P., Soler-Vidal, J., Ramiro-Sousa, N., Rodriguez-Martinez, A., Sarri-Closa, C., Sarró, S., Larrubia, J., Andrés-Bergareche, H., Miguel-Cesma, M.C., and Padilla, P.P., "The linguistic signature of hallucinated voice talk in schizophrenia", Schizophrenia Res., vol. 206, pp. 111–117, 2019.

[34] Passchier-Vermeer ,W. and Passchier, W.F., "Noise exposure and public health", Environ. Health Perspect., vol. 108(1), pp. 123–131, 2000.

[35] Stansfeld, S.A. and Matheson, M.P., "Noise pollution: non-auditory effects on health", British Medical Bulletin, vol. 68(1), pp. 243–257, 2003.

[36] Letowski, T.R. and Scharine, A.A., "Correctional analysis of speech intelligibility tests and metrics for speech transmission", report, U.S. Army Research Laboratory, pp. 1–50, 2017.

[37] 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.

Read Speech Protocol Criteria for Speech-Based Health Screening Applications

Brian Stasak[1,2] and Julien Epps[2]

[1]Black Dog Institute, Sydney, NSW – Australia
[2]School of Elec. Eng. & Telecom., UNSW Sydney, NSW – Australia
b.stasak@unsw.edu.au, j.epps@unsw.edu.au

Abstract

Automatic speech-based processing using machine learning is expanding in digital healthcare, bolstering potential as a non-invasive, remote medical screening tool. There is currently a need for deeper understanding of read speech protocols and applying systematic measurements to help tailor new protocols with deliberate attribute criteria. This study investigates eight read speech protocols commonly utilized to study speech behaviours and proposes two new protocols with greater criteria extremes. An investigation of text, phonetic, linguistic, and affective proposed criteria automatically extracted from these read speech protocols reveals important merits and limitations for use in speech-based digital health screening applications.
Index Terms: data collection; elicitation; voice processing

1. Introduction

Healthcare clinicians perform speech-language evaluations to screen, diagnose, and monitor many disorders [1, 2]. During a structured interview, clinicians observe the patient's speech production, such as articulation, breathing, phonation, and voice quality. Clinicians also evaluate a patient's spoken language ability, including grammar, pragmatics, memory, and expressive capacity. Abnormal speech-language symptoms are often early precursors to a variety of disorders and illnesses.

Some clinical evaluations also include an analysis of speech behaviours based on read protocols, which consists of pre-selected words, sentences, or paragraphs that are read aloud. Advantages to read speech protocols include minimal instruction, repeatability, 'ground-truth' knowledge, limited cognitive scope, and controlled phonetic variability. Further, read speech protocols are relatively easy to implement in digital smart device apps. Among the most popular read speech protocols for speech-based analysis are: 'Arthur' [3], 'The Caterpillar' [4], 'The Farm Script' [5], 'Hunter Script' [5], 'The Grandfather Passage' [6], 'The John Passage' [7], 'The North Wind and the Sun' [8], and 'The Rainbow Passage' [9].[1]

When selecting an ideal read speech protocol for pathological medical analysis there are important factors, such as speaker background (e.g., age, reading skill level), illness specificity (i.e., focus on elicitation of key symptoms), and task duration (i.e., time duration, number of samples needed). Despite read speech protocol groundwork [6, 9-12], still many speech-based health screening studies [13-15] continue to use semi-antiquated read speech protocols (i.e., the origin of 'Arthur', 'The Grandfather Passage', and 'The North Wind and the Sun' were derived from writings more than a century old), which were not originally intended for 'universal' illness/disorder screening. Of the protocols mentioned, many were originally used to subjectively judge verbal intelligibility oral reading rates of school-aged children and not designed for precision assessment of a wide variety of different types of disorders (e.g., psychogenic, neurological, respiratory, voice) [3, 6, 8, 9]. For example, 'The Rainbow Passage' is nearly 80 years old, and it was based on child voice articulation drills described for '*correction of disorders*'; although it does not define which ones or provide statistical normative data [9].

While it is frequently believed that these read speech protocols contain every English phoneme and are phonetically balanced [16], previous studies [11, 17, 18] have shown many of these read speech protocols have non-ideal, disproportionate phoneme distribution ratios. Moreover, some of these protocols do not purposefully isolate or repeat specific phonemes, phonetic transitions, and/or language components that may be more useful in more direct screening for certain illnesses/disorders. Increasing the opportunity for phonemes in read speech protocols that ordinarily are less frequently found in conversational speech may be advantageous, especially if a disorder (e.g., aphasia, apraxia, dysphonia) affects production of a specific phoneme or phoneme class.

From an age-appropriate readability standpoint, many of these read speech protocols [3, 7, 8, 9] are third-person narratives that include unfamiliar themes and advanced vocabulary (i.e., 'The North Wind and the Sun' is a translation based on Aesop's ancient Greek fable [8]). This may add more difficulty reading aloud when compared with natural conversational speech. Also, some of these read speech protocols include unsuitable themes for younger children (i.e., 'Arthur' has a death; 'The Hunter' has a firearm).

Previous studies [4, 16-18, 20] have comparatively examined read speech protocols. However, these previous studies focused primarily on text-based attributes or phonetic distributions rather than deeper level structure criteria, such as age of articulatory mastery and gestural phonetic transition effort. Only two of these previous studies [4, 20] contributed new read speech protocols – working to further expand the acoustic speech elicitation space.

Automatic digital speech-based screening studies [4, 13, 21-26] involving the aforementioned protocols typically comprised a relatively limited number of speakers (i.e., less than two dozen) and each contained dissimilar speakers (e.g., children, adults, elderly). Also, these automatic speech-based illness screening studies examined the effectiveness of just one or two read speech protocols. Further, only recently have speech-based illness detection studies [14, 27-30] examined linguistic/affect components in read speech protocols. For example, Boaz et al. [27] investigated 'The Rainbow Passage'

[1] Free access to the speech protocols in this study are available at: https://www.researchgate.net/profile/Brian_Stasak/projects/

and how its affective content impacted speech disfluencies. They found 'The Rainbow Passage' and a novel constructed passage disfluency types (e.g., pauses, repeats) were not significantly different. However, half of the individuals studied showed a large difference in the number of disfluencies between the two passages, likely due to readability factors.

This paper investigates text, phonological, linguistic, and affective attributes from ten read speech protocols, including two new speech protocols. New articulatory criteria regarding age of acquisition mastery, gestural effort, and phoneme-to-word ratios are reported along with examples of how different criteria are relevant to disorders. Results herein show that some read speech protocols are more alike than others. This deeper level analysis adds to the discussion on the need for modern read speech protocols with more methodical design criteria, including phonetic, linguistic, and affect factors, resulting in greater knowledge of how age- and skill-appropriate a protocol is for a particular speech-based medical screening.

2. Analysis and Discussion

2.1. Protocols and Criteria System

Eight existing read speech protocols were examined in this study (see Tables 1 and 2). These established protocols were selected for analysis due to their frequent use in previous speech-based studies [4, 13, 16-18, 20-26]. Two newly proposed read speech protocols, 'Jazz' and 'Restaurant', were created using deliberately chosen words to induce higher articulatory skill level and unique read token word demand.

In Figure 1, a system design for automatic extraction of criteria based on existing or new speech protocol texts is proposed. Using raw text as input, extraction of various feature types produces an output, whereby feature values are then compared to other read texts. This allows an understanding of deeper level information about what a read protocol contains. A set of feature criteria ranges can be experimented with to help better tailor read speech protocols for more specific precision disorder screening. This proposed system also can be utilized for automatic data selection of non-read free speech transcripts to extract single phrases with feature criteria ranges of interest.

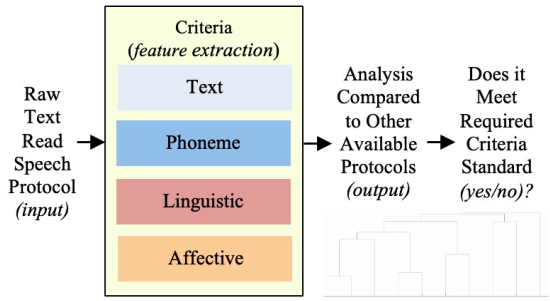


Figure 1: Proposed multi-dimensional automatic read speech protocol analysis design, which provides in-depth knowledge of different criteria, important when considering the appropriateness of read speech protocol for illness screening applications and test subject demographics. A hierarchical cluster tree method (e.g., Chebychev, Spearman) can be used to calculate read speech protocol similarity to other protocols [17].

2.2. Text and Phoneme Criteria

Surface-level text-based analysis provides limited articulatory insight. For example, a six-letter word like '*though*' only contains two phonemes, whereas a six-letter word like '*plants*' contains six unique phonemes. Therefore, text-based letter/word counts can be misleading in terms of articulatory ground-truth (e.g., *fizz*, *physics*, *laugh*), phoneme transition articulatory demand, and phoneme distributions.

The text-based analysis of the read speech protocols is shown in Table 1. Per protocol, the total number of words varied from as low as 97 to as high as 338. The mean length utterance (i.e., average number of words per sentence) of a read speech protocol, also referred to as MLU, may be an important consideration for automatic speech-based illness screening applications because recorded read aloud tasks that require minutes to complete require more time, storage, and user commitment/focus. Table 1 shows that the 'The North Wind

Table 1: Read speech protocol comparison of basic text and phonetic information. This also includes never-before reported protocol scores for articulatory acquisition mastery and gestural effort measures [28, 29]. Newly proposed read speech protocols are shaded.

Protocol	# Words	# Unique Words	# Sent.	MLU	Word Range per Sent.	Aver. TTR	# Phonemes	MLP	Phon. # Range per Sentence	Aver. P/W Ratio	Aver. Art. Mast.	Aver. Gest. Eff.
Arthur	338	197	30	11.27	1 – 24	0.58	1023	33.1	4 – 74	3.12	62	6.3
Caterpillar	195	115	16	12.19	4 – 26	0.59	664	41.5	9 – 77	3.37	61	6.5
Farm	313	169	16	19.56	5 – 34	0.54	902	56.3	16 – 101	2.87	62	6.6
Grandfather	131	99	7	16.63	8 – 29	0.76	475	59.4	27 – 116	3.59	62	6.3
Hunter	279	151	17	16.41	4 – 35	0.54	878	51.7	13 – 109	3.21	62	6.4
John	191	119	11	17.36	7 – 29	0.62	617	56.0	25 – 91	3.25	58	6.8
North	113	64	5	22.60	14 – 36	0.57	384	76.8	52 – 115	3.44	62	6.5
Rainbow	330	175	19	17.37	8 – 36	0.53	1146	60.3	26 – 117	3.49	62	6.4
Jazz	117	88	12	9.75	5 – 16	0.75	478	39.8	21 – 64	4.15	67	6.3
Restaurant	97	76	9	10.78	5 – 14	0.78	446	49.6	25 – 60	4.68	65	6.7

Table 2: Read speech protocol comparison of basic linguistic and affective information. This includes Flesch-Kincaid grade-level and reading ease, Dirichlet allocation age of word exposure and acquisition scores [4, 31, 32]. Affect averages are also shown derived from Affective Norms for English Words (ANEW) [33]. All affective scores had a range from 1 (low) to 9 (high).

Protocol	% Passive Voice	Aver. Grade Level	Aver. Reading Ease	Aver. Age Word Exposure	Aver. Age Word Acquisition	Arousal	Dominance	Valence
Arthur	0.0	1.9	100	3.0	4.98	4.99	5.10	4.93
Caterpillar	0.0	5.0	81	3.7	5.07	5.33	5.29	6.26
Farm	18.7	2.6	100	2.5	4.86	4.72	5.51	6.22
Grandfather	0.0	5.2	81	3.5	5.77	4.89	5.66	6.53
Hunter	0.0	3.9	96	3.3	5.17	4.86	4.96	6.06
John	9.0	6.4	80	4.0	5.64	4.63	5.53	6.24
North	40.0	8.2	76	4.5	5.55	5.43	6.40	7.03
Rainbow	15.7	7.7	70	4.4	5.48	5.07	5.33	8.86
Jazz	8.3	6.6	64	6.4	6.58	5.33	5.40	6.52
Restaurant	0.0	9.5	46	5.0	6.30	5.03	5.69	6.66

and the Sun' protocol has the largest average for number of words per sentence (22.6), whereas the new 'Jazz' protocol averages much less (9.75).

The number of unique words was much less than the total number of words for all protocols (76–197). A type-token-ratio (TTR) is defined as the relationship between the number of unique words (e.g., core word types excluding word modifiers) and the number of total words (e.g., tokens). It is calculated by dividing the number of unique words by the total number of words. The more unique words there are in comparison to the number of total words, then the more lexical variety (i.e., a high TTR). Previously in Table 1, it was reported that the 'Restaurant' protocol had the highest TTR, whereas the 'The Rainbow Passage' demonstrated the lowest TTR.

Generally, the more words that are repeated, the more opportunity for like-word acoustic comparison. Therefore, depending on the illness/disorder being screened, it may be more effective to use a read speech protocol that has a low TTR to evaluate the same words more than once. On the contrary, if lexical diversity is of higher interest, it may be optimal to utilize a read speech protocol that has a high TTR. For example, for an early childhood speech sound disorder, having multiple examples of the same word may be good to determine the percentage of correct pronunciation, whereas for short-term memory disorders a read aloud memory test, a read speech protocol with more unique words helps, to increase cognitive demand and test specific keyword recall.

A phoneme-based analysis of the read speech protocols was conducted using thirty-nine English phonemes based on the Carnegie Mellon University phonetic dictionary [34]. A python script was created to convert the raw text of each word per read protocol into the most common phonetic representation. Results demonstrated that none of the read protocols were truly phonetically balanced; meaning they did not have equal representations for each English phoneme. Therefore, using these protocols, it may be difficult to analyse multiple examples of specific phonemes (e.g., /j, ʒ/) when compared with other phonemes that have a much greater frequency (e.g., /ə, n/). For example, utilizing a protocol with a higher frequency of rarer phonemes is important if close analysis of palatal or postalveolar positioned articulatory production is of high interest (e.g., speech sound disorders, palatal fronting).

In comparing each of the read speech protocols to standard norm English phoneme distributions found in natural free conversational speech [35], it was observed that the read speech protocols had a very strong 0.83-0.91 Spearman's rank correlation coefficient (a = 0.05). This analysis indicates that these read speech protocols contain similar phoneme distributions to natural speech. Phonemic analysis per read speech protocol showed large differences in the number of English phonemes, especially mean length of sentence phonemes (MLP) (i.e., the average number of phonemes per sentence). As shown previously in Table 1, although the 'Arthur' (1023) and 'The Rainbow Passage' (1146) contained the largest number of phonemes, 'The North Wind and the Sun' protocol had the largest average number of phonemes per sentence (~77) despite roughly one-third fewer sentences.

The phoneme-to-word ratio (P/W) is a better indicator of articulatory demand than using an average text-based word length or word total because it represents a ground-truth of spoken sounds. Further, read speech protocols with higher density phoneme-to-word ratios enable generation of more phonemes using fewer words – therefore, allowing increased efficiency in comparison to longer protocols that contain more words and recording time.

In terms of the phoneme-to-word ratio, which was calculated by dividing the number of phonemes by the total number of words, Table 1 and Figure 2 show that the 'Jazz' and 'Restaurant' new read speech protocols were much higher than the other traditional protocols. Analysis based on sentence-level average phoneme-to-word ratio scores indicated that the 'Jazz', 'Restaurant', and 'The Farm Script' were least like the other read speech protocols. The 'Hunter', 'John', 'Arthur', 'Caterpillar', and 'North Wind' were most alike in terms of phoneme-to-word ratio scores.

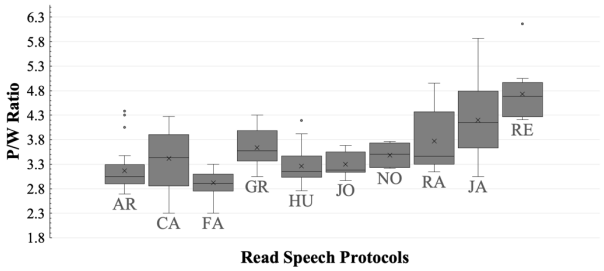


Figure 2: Sentence-level distributions of phoneme-to-word ratio (P/W) per read speech protocol: Arthur (AR), Caterpillar (CA), Farmer (FA), Grandfather (GR), Hunter (Hu), John (JO), North (NO), Rainbow (RA), Jazz (JA), and Restaurant (RE). Mean is indicated by 'x', the solid line indicates median, and outliers are represented as dots.

Previously, an acoustic speech-based study [28] showed that phrases containing a greater number of consonant phonemes mastered later in life were more effective in detecting individuals with depression. The reasoning is that phonemes mastered later in life require greater articulatory coordination, and are therefore a better measure for fine motor control. Shown in Figure 3, an assessment of age of articulatory acquisition mastery demonstrated that the 'Jazz' protocol had the highest average age in months (67), whereas the 'John' protocol had the lowest average (58). The new read speech protocols had an average articulatory age of acquisition mastery that was higher than the other existing protocols. The 'Arthur', 'Grandfather', 'Hunter', 'Rainbow', and 'North Wind' are most alike based on mean age of articulatory acquisition mastery.

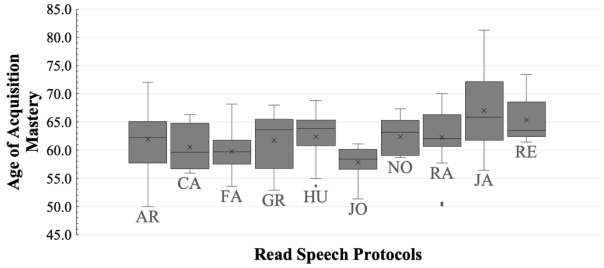


Figure 3: Sentence-level distributions of age of acquisition mastery per read speech protocol: Arthur (AR), Caterpillar (CA), Farmer (FA), Grandfather (GR), Hunter (HU), John (JO), North (NO), Rainbow (RA), Jazz (JA), and Restaurant (RE). The age of acquisition mastery scale is in number of months.

Similarly to [29], gestural effort, which is the amount of articulatory change within an utterance, was examined for the ten read speech protocols. The gestural measure was used to analysed seventeen different Chomsky-Halle articulatory manners using binary representations, whereby the greater number of switches between phoneme manners resulted in a higher gestural effort value (i.e., Hamming distance). It is believed that more rapid manner activation/non-activation productions between each phoneme necessitates increased fine

motor control (e.g., motoric coordination). Moreover, [29] found that recorded utterances with a higher gestural effort measure produced improved automatic depression detection. In Figure 4, gestural effort analysis demonstrated that 'Arthur' was the broadest in terms of individual sentence ranges. Based on mean gestural effort, the 'John' and 'Restaurant' read speech protocols were the most demanding.

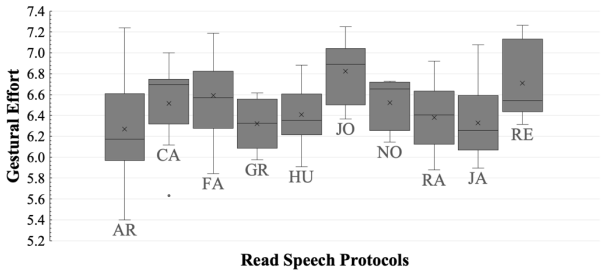


Figure 4: Sentence-level distributions of gestural effort per read speech protocol: Arthur (AR), Caterpillar (CA), Farmer (FA), Grandfather (GR), Hunter (HU), John (JO), Northwind (NO), Rainbow (RA), Jazz (JA), and Restaurant (RE).

It is mostly unknown what kind of impact multi-dimensional criteria aspects have on automatic speech-based illness detection. But future investigation of speech-based digital health screening applications using read speech protocol may reveal improved screening capabilities, especially new protocols tailored towards specific illnesses/disorders. While it is known that criteria (e.g., phonetic, linguistic, affective) influence each other and the acoustic speech signal, there is no known protocol that simultaneously covers the entire ranges of these criteria. Figure 5 shows that even within considering three phoneme-based criteria, read speech protocol ranges can vary. For example, in Figure 5, it is observed that the 'Arthur' protocol is more suitable for younger test subjects, wherein a longer and a more repetitive token word sample is required (i.e., but it has unsuitable death theme). On the contrary, the 'Restaurant' protocol is better suited for older test subjects, wherein a shorter and a less repetitive token word sample is desirable. Knowledge regarding criteria restraints within read speech protocols further provides better insight towards which specific read speech protocol to select for health screening purposes.

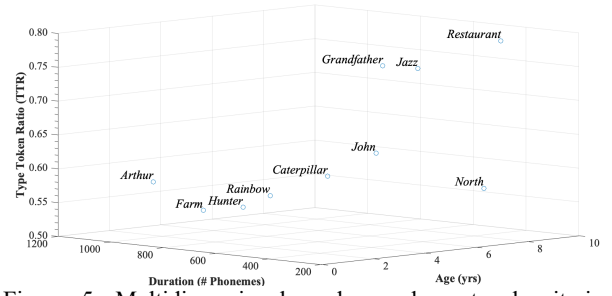


Figure 5: Multidimensional read speech protocol criteria. Depending on speaker demographic and digital health criteria needs, careful consideration should be taken to make sure protocols are appropriate for digital health applications.

2.3. Linguistic and Affective Criteria

The 'Farm', 'John', 'North Wind', 'Rainbow', and 'Jazz' read speech protocols include some sentences that contained passive voice grammar. However, the passive voice maximum for all protocols examined was below 40%, indicating that new protocols could investigate greater use of passive voice. For example, it is well-known that individuals with dementia usually exhibit poor cognitive function, word retrieval difficulty, increased speech disfluencies, and struggle with constructing or recalling novel appropriate responses, especially from passive voice texts [36].

The reading level for the read ten speech protocols were calculated using the Flesch-Kincaid method for average grade level and average reading ease [31]. Analysis in Table 2 indicated that the 'Restaurant' (9.5) and 'North Wind' (8.2) read speech protocols may be inappropriate for evaluating speech of individuals younger than eighth grade based on reading ease scores. Strangely, the reading ease score for the 'Rainbow Passage' (7.7) somewhat contradicts [9], where the introduction suggests its broad use for '*school-aged children*'.

Because reading is a requirement for these protocols, linguistic norm criteria should be closely examined to establish age appropriateness and/or level of difficulty [3, 31, 32]. In Table 2, the average age of word exposure for 'Jazz' was the highest (6.4) due to many less frequent keywords. The 'Jazz' and 'Restaurant' read protocols also demonstrated the highest values for average word acquisition (6.58, 6.30), whereas 'Farm' contained average word acquisition that was much lower (4.86). Table 2 analysis demonstrates that word grade level, reading ease, exposure, and acquisition are different criteria that do not always directly associate with each other.

In Table 2, using affective norms text-processing software [33], an examination of affect across the ten different read speech protocols reveals that the degree of arousal is narrow and typically in the neutral range. Dominance was also shown to be typically neutral for the read speech protocols, except for the 'North Wind', which had greater dominance (6.40). Moreover, valence average per read speech protocol demonstrated a bias towards positive valence keywords. The only exception to this was the 'Arthur' read speech protocol that had a neutral valence value of 4.93, nearing the negative valence range (≤ 4.50). Affective results herein show that there is a need for read speech protocols where arousal, dominance, and valence are more extreme to understand paralinguistic behaviours in different emotional contexts. A recent speech-based study [30] on automatic mood disorder detection indicated that protocols containing a broader range of valence can help improve detection of individuals with mental illness.

3. Conclusion

It is vital that a greater number of criteria are taken into account than most current speech-based studies when designing, choosing, and executing speech elicitation materials for digital health analysis. It is likely that deliberately tailored protocol design for speech-based digital health applications will produce greater benefit, with less computation delving through unnecessary excess acoustic speech data to analyze only a small percentage. Also, protocol designs may have advantages over free speech collections because they can potentially isolate speech and/or language tasks that might otherwise involve a mixture of cognitive, memory, and motor articulation skill level. Some isolated read verbal tasks already exist, such as the diadochokinetic and Stroop color test protocols, but their speech focus is more obvious to the participant, repetitive, and less natural – which may influence speech behaviors.

Generally, there is much greater room for new development in the field of speech-language elicitation protocols for health screening purposes. Future studies on illness/disorders should directly measure many different read speech protocol sensitivity/specificity results to help determine which are the best depending on the illness of interest and test subject age.

4. References

[1] Chevrie-Muller, C., Sevestre, P., & Seguier, N., "Speech and psychopathology", Lang. and Speech, vol. 28 (1), pp. 57-79, 1985.

[2] Hirschberg, J., Hjalmarsson, A., & Elhadad, N., "You're as sick as you sound: using computational approaches for modeling speaker state to gauge illness and recovery", In: A. Neustein (ed.) Advances in Speech Recognition: Mobile Environments, Call Centers and Clinics, Springer Science + Business Media, pp. 305-322, 2010.

[3] MacMahon, M.K.C., "The woman behind 'Arthur'", J. Intern. Phonetic Assoc., vol. 21 (1), pp. 29-31, 1991.

[4] Patel, R., Connaghan, K., Franco, D., Edsall, E., Forgit, D., Olsen, L., Ramage, L., Tyler, E., & Russell, S., "The caterpillar: a novel reading passage for assessment of motor speech disorders", Am. J. Spch. Lang. Path., vol. 22 (1), pp. 1-9, 2013.

[5] Crystal, T.H., & House, A.S., "Segmental duration in connected speech signals: preliminary results", J. Acoust. Soc. Am., vol. 72 (3), pp. 705-716, 1982.

[6] Van Riper, C., Speech Correction (4th Ed.), Prentice Hall, Englewood Cliffs, NJ – USA, 1963.

[7] Tjaden, K., & Wilding, G., "Rate and loudness manipulations in dysarthria: acoustic and perceptual findings", J. Speech Lang. Hear. Res., vol. 47 (4), pp. 766-783, 2004.

[8] Townsend, G.F., Aesop's Fables, George Routledge & Sons, London & New York, 1868.

[9] Fairbanks, G., Voice and Articulation Drillbook (2nd Ed.), Harper & Row, New York – USA, pp. 124-139, 1960.

[10] Egan, J.P., "Articulation testing methods", Laryngoscope, vol. 58, pp. 955-991, 1948.

[11] Patel, R.R., Awan, S.N., Barkmeier-Kraemer, J., Courey, M., Deliyski, D., Eadie, T., Paul, D., Svec, J., and Hillman, R., "Recommended protocols for instrumental assessment of voice: American speech-language-hearing association expert panel to develop a protocol for instrumental assessment of vocal function", American J. Speech-Language Pathology, vol. 27 (3), pp. 887-905, 2018.

[12] Schiel, F., Draxler, C., Baumann, A., Ellbogen, T., & Steffen, A., "The production of speech corpora", Version 2.5, Bavarian Arch. for Speech Signals, University of Munich, 2004.

[13] Mundt, J.C., Snyder, P.J., Cannizzaro, M.S., Chappie, K., and Geralts, D.S., "Voice acoustic measure of depression severity and treatment response collected via interactive voice response (IVR) technology", J. Neurolinguistics, vol. 20 (1), pp. 50-64, 2011.

[14] Jaing, H., Hu, B., Liu, Z., Lihua, Y., Wang, T., Liu, F., Kang, H., and Li, X., "Investigation of different speech types and emotions for detecting depression using different classifiers, Speech Comm., vol. 90, pp. 39-46, 2017.

[15] Williamson, J.R., Quatieri, T.F., Helfer, B.S., Horwitz, R., Yu, B., and Mehta, D.D., "Vocal biomarkers of depression based on motor incoordination", In: Proc. AVEC '13: Proceedings of the 3rd ACM international workshop on Audio/visual emotion challenge, pp. 41-48, 2013.

[16] Ludlow, C.L., Kent, R.D., and Gray, L.C., Measuring Voice, Speech, and Swallowing in the clinic and Laboratory, Plural Publishing Inc., San Diego, CA – USA, 2019.

[17] Lammert, A.C., Melot, J., Sturim, D.E., Hannon, D.J., DeLaura, R., Williamson, J.R., Ciccarelli, G., & Quatieri, T.F., "Analysis of phonetic balance in standard English passages", J. Speech, Lang., and Hearing Res., vol. 63 (4), pp. 917-930, 2020.

[18] Powell, T.W., "A comparison of English reading passages for elicitation of speech samples from clinical populations", Clinical Linguistic & Phonetics, vol. 20, pp. 91-97, 2006.

[19] Kent, R.D., Kent, J.F., & Rosenbek, J.C., "Maximum performance tests of speech production", J. of Speech Hear. Disord., vol. 52, pp. 367-387, 1987.

[20] Deterding, D., "The north wind versus a wolf: short text for the description and measurement of English pronunciation", J. Intern. Phonetic Assoc., vol. 36 (2), pp. 187-196, 2006.

[21] Bayestehtashk, A., Asgari, M., Shafran, I., and McNames, J., "Fully automated assessment of the severity of Parkinson's disease from speech", Comp. Speech Lang., vol. 29 (1), pp. 172-185, 2015.

[22] Perez, M., Jin, W., Le, D., Carlozzi, N., Dayalu, P., Roberts, A., and Mower-Provost, E., "Classification of Huntington disease using acoustic and lexical features", In: Proc. INTERSPEECH 2018, pp. 1898-1902, 2018.

[23] Turner, G.S., Tjaden, K., and Weismer, G., "The influence of speaking rate on vowel space and speech intelligibility for individuals with amyotrophic lateral sclerosis", J. Speech, Lang., and Hearing Res., vol. 38 (5), pp. 1001-1013, 1995.

[24] Whitfield, J.A., Kriegel, Z., FullenKamp, A.M., and Mehta, D.D., "Effects of concurrent manual task performance on connected speech acoustics in individuals with Parkinson disease", J. Speech, Lang., Hearing Res., vol. 62 (7), pp. 2099-2117, 2019.

[25] Howell, P., Sackin, S., and Glenn, K., "Development of a two-stage procedure for the automatic recognition of dysfluencies in the speech of children who stutter: II. ANN recognition of repetitions and prolongations with supplied word segment markers", J. Speech, Lang., Hearing Res., vol. 40, pp. 1085-1096, 1997.

[26] Yunusova, Y., Weismer, G., Kent, R.D., and Rusche, N.M., "Breath-group intelligibility in dysarthria: characteristics and underlying correlates", J. Speech, Lang., and Hearing Res., vol. 48, pp. 1294-1310, 2005.

[27] Ben-David, B.M., Moral, M., Namasivayam, A., & van Lieshout, P., "Linguistic and emotional-valence characteristics of reading passages for clinical use and research", J. of Fluency Disord., vol. 49, pp. 1-12, 2016.

[28] Stasak, B., Epps, J., & Goecke, R., "Elicitation design for acoustic depression classification: an investigation of articulation effort, linguistic complexity, and word affect", INTERSPEECH '17, Stockholm – Sweden, pp. 834-838, 2017.

[29] Stasak, B., Epps, J., and Lawson, A., "Analysis of phonetic markedness and gestural effort measures for acoustic speech-based depression classification", In: Proc. 2017 Seventh International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW), San Antonio, TX – USA, pp. 165-170, 2017.

[30] Stasak, B., and Epps, J., "Automatic depression classification based on affective read sentences: opportunities for text-dependent analysis", *Speech Comm.*, vol. 115, pp. 1-14, 2019.

[31] Kincaid, J.P., Fishburne, R.P., Rogers, R.L., and Chissom, B.S., "Derivation of new readability formulas (automated readability index, fog count, and Flesch reading ease formula) for Navy enlisted personnel", U.S. Naval Research Branch Report, pp. 8-75.

[32] Kyle, K. & Crossley, S.A., "Automatically assessing lexical sophistication: Indices, tools, findings, and application", TESOL Quarterly, vol. 49 (4), pp. 757-786, 2015.

[33] Crossley, S.A., Kyle, K., & McNamara, D.S., "Sentiment analysis and social cognition engine (SEANCE): An automatic tool for sentiment, social cognition, and social order analysis", Behavior Research Methods, vol. 49 (3), pp. 803-821, 2017.

[34] Carnegie Mellon University (CMU), 1993. The Carnegie Mellon Pronouncing Dictionary v0.1. Carnegie Mellon University: http://www.speech.cs.cmu.edu/cgi- bin/cmudict

[35] Mines, M.A., Hanson, B.F., and Shoup, J.E., "Frequency of occurrence of phonemes in conversational English", Language and Speech, vol. 21 (3), pp. 221-241, 1978.

[36] Emery, V.O.B., "Language impairment in dementia of the Alzheimer type: a hierarchical decline?", Inter. J. Psych. Medicine, vol. 30 (2), pp. 145-164, 2000.

A semi-automatic workflow for orthographic transcription of a novel speech corpus: A case study of AusKidTalk

Tünde Szalay[1,2], Louise Ratko[3], Mostafa Shahin[2], Tharmakulasingam Sirojan[2], Kirrie Ballard[1], Felicity Cox[3], Beena Ahmed[2]

[1]The University of Sydney, [2]The University of New South Wales, [3] Macquarie University

tuende.szalay@sydney.edu.au

Abstract

Developing automatic speech recognition (ASR) tools for AusKidTalk, the novel Australian English (AusE) children's corpus, presents a circular problem: AusKidTalk is designed to develop adequate ASR for AusE-speaking children; however, orthographic transcription of AusKidTalk requires ASR tools not yet developed. Our semi-automatic workflow augments existing (but inadequate) automatic tools with manual transcription. IBM-Watson diarisation and UNSW ASR orthographic transcription automatically generate Praat textgrids with time-aligned orthographic transcriptions. A webtool distributes the textgrids, collects manual corrections, and implements consistency checks. Manual correction is conducted with a custom Praat interface. The output is a searchable, orthographically transcribed, and time-aligned corpus.

Index Terms: audio corpus, orthographic transcription, automatic speech recognition for novel populations

1. Introduction

AusKidTalk is an audio-visual corpus of Australian English (AusE) speaking children [1, 2]. Orthographic transcription and annotation of AusKidTalk are the essential first steps towards obtaining phoneme-level annotation [2, 3]. Due to its size, cost-efficient transcription and annotation is only possible with automatic speech recognition (ASR) tools.

Current ASR systems have been developed for adult speech and their performance drops considerably on children's data due to developmental differences [4, 5]. As ASR systems are trained on vast amounts of domain-specific annotated speech data [6], developing ASR for children has been thwarted by the lack of available children's corpora. Currently only 15 children's speech corpora are publicly available worldwide [7]. All were collected using problem-specific protocols with limited tasks, none is fully annotated, and only three of them are sufficiently sized for ASR development [7]. Developing new ASR tools for a novel large corpus presents a circular problem: one of the aims of AusKidTalk is to develop new and accurate ASR for AusE-speaking children [2]; but to efficiently annotate AusKidTalk, accurate ASR tools not yet developed are required.

To overcome the lack of suitable ASR tools, we developed a multi-step workflow combining existing, but suboptimal ASR tools designed for other populations, augmented with manual correction, to provide orthographic annotation for parts of AusKidTalk. Our aim was to balance the efficiency of existing ASR tools with the accuracy of manual annotation. This paper is a case study in corpus building, describing the challenges of orthographically transcribing AusKidTalk and presenting our solutions through a step-by-step guide of our workflow.

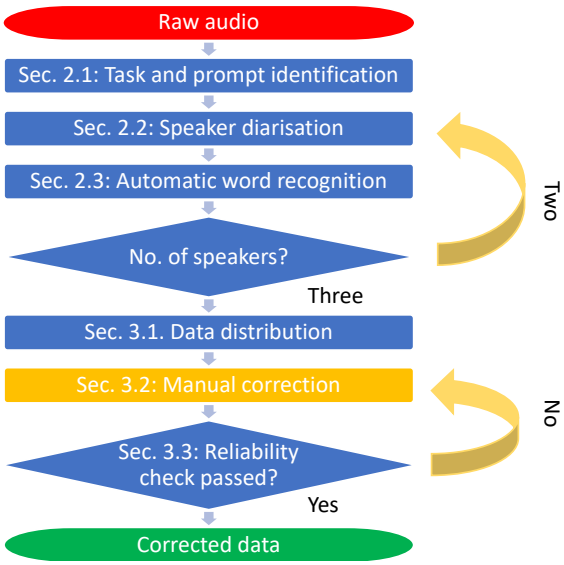


Figure 1: *Workflow outline. Red: start. Green: end. Blue: automated process. Yellow: manual process. Diamond: decision.*

1.1. The AusKidTalk corpus

To create the AusKidTalk corpus, we are collecting data from 750 native speakers of AusE, aged 3–12, with and without speech disorders and autism spectrum disorder, who contribute 90-120 minutes of audio. 475 children have participated; data collection is ongoing. Children complete five tasks with a range of linguistic complexity: three are prompted (word elicitation, pseudoword- and sentence repetition) and two are spontaneous (story telling and emotion elicitation); for details, see [2].

Tasks are presented via an Android app on a tablet while speech is recorded onto a PC. There is no direct synchronisation between the tasks, the prompts that appear on the tablet, and the recorded audio file. To align speech, the Android app plays a 1s high-frequency tone at the start of each task and records timestamps at the start and end of each task and prompt.

1.2. Challenges in annotating children's data

The audio recordings contain varied, spontaneous, and unexpected speech, inherent to children's data. There are unexpected responses to the prompts, such as responding to a picture of a cucumber with "zucchini", with a giggle to a picture of a belly-button, or only repeating parts of a sentence. In all tasks, there are non-task-related conversations between the child and the interviewer leading the recording session.

As the entire conversation was recorded due to the child's headset microphone picking up all speakers, the audio contains three speakers: the child, the pre-recorded model speaker who produced verbal prompts for the word and sentence level tasks, and the interviewer, instructing and aiding the child (e.g., "Can you speak up a bit?", "Very good!"). The combination of unexpected responses, spontaneous conversations, and three distinct speakers results in a high volume of non-target speech which further increases the difficulty of automated annotation.

1.3. Scope and goal of the annotation workflow

The goal of the workflow protocol described here (Fig. 1) is to annotate the prompted picture naming task (Task 1) for each child by orthographically transcribing all 130 target words and locating their start and end times in the audio file. Our output is a time-aligned Praat textgrid for each child that contains intervals for the target words only and an easily searchable csv file listing the target words with their start and end times.

2. Automated tools

2.1. Time-aligning with tone detection

To determine the start and end of the word elicitation task in the audio file, we developed automated tone detection using a non-linear binary Support Vector Machine (SVM) with the radial basis function kernel to identify the location of the 1s tone and match it with the timestamps. To train the SVM, feature vectors were extracted from 3700 not-tone and 3700 tone frames selected randomly from 10 recordings and spliced with the feature vector of two preceding and two succeeding frames. On a test set of 10 recordings, our classifier achieved 0% False Acceptance Rate by never identifying not-tone as tone and approximately 10% False Rejection Rate by missing 4 out of 45 tones.

The SVM classifies each 10ms frame of the recording as tone or not-tone. The moving average of the number of detected tones is calculated using a one-second sliding window. Peak points with a moving average above 0.9, i.e., with at least 90% of frames classified as tone within a 1s window, are considered to be tone positions. The duration between every two tone sounds is calculated and compared to the duration between every two timestamps marking the start and the end of a task. Reference tone-timestamp pairs are identified when the duration between any two tone sounds is equal to the duration between any two timestamps. Task 1 is separated in the audio file using reference tone-timestamps.

A Praat textgrid is generated with intervals indicating the start and end of each prompt using the prompt timestamps. The prompt interval is the time between the presentation of the prompt picture to the child, and the pressing of the assessment button by the interviewer which indicates that the child completed the attempt of the current prompt.

2.2. Diarisation with IBM-Watson

To separate the child's voice from the interviewer and the model speaker, the IBM-Watson speech-to-text web service is used to map the three speakers onto three different tiers on a Praat textgrid. IBM-Watson uses deep learning AI with language specific models of grammar, vocabulary and acoustics to diarise and transcribe speech. The "AU-Narrow Band" model has been tested, as it is suitable for audio sampling rates above 8 kHz and can diarise and transcribe the speech of up to three speakers. The "AU- Narrow Band" model will be phased out in 2023, and we will replace it with the "AU Multimedia model"; the current paper reports data using the former.

Task 1 audio files are resampled from 44.1 kHz to 12 kHz to reduce uploading time and processed with IBM-Watson. The resulting JSON files contain utterance labels, confidence scores, and start and end times. Transcription and diarisation information is converted into Praat textgrids, with one speaker per tier.

IBM-Watson may diarise the audio correctly as containing three speakers or incorrectly as containing two speakers. To ensure that the child's speech is identified, tiers are counted automatically. Textgrids containing fewer than three tiers per recording are visually inspected by a trained phonetician to identify which tier does not contain the child's speech (top yellow arrow in Fig. 1). When the child's tier is separate from the other two speakers, the audio and the textgrid is passed onto the UNSW ASR tool (Sec. 2.3). When the child's speech is not separate from either the interviewer or the model speaker, speech on the non-child tier is silenced using a Praat script that identifies the intervals on the non-child tier and reduces their amplitude to zero. IBM-Watson is redeployed to differentiate the child's speech from the remaining speaker. A second JSON file is returned and converted into a Praat textgrid, with two tiers, one for the child, and one for the other speaker.

In a test sample of five randomly selected recordings, IBM-Watson correctly identified the three different voices in the file in two recordings. The remaining three were diarised as two speakers; having silenced the non-child audio and redeploying IBM-Watson resulted in a total of three speakers. That is, IBM-Watson successfully identified three distinct speakers in five out of five recordings. Diarisation and transcription accuracy of the tier identified as child only was evaluated in test session recordings from four children (age range = 4 – 10 years, mean = 6.75). 82%–95% (mean = 89%) of all intervals mapped onto the child's tier contained the child's speech. 85%–91% (mean = 87.93%) of all targets were identified on the child's tier, while the remaining targets were mapped onto another speaker's tier. That is, the child's speech was separated from the other speakers' with high accuracy. Orthographic transcription accuracy however was so low as to being practically unusable with a word error rate ranging from 94% to 57% per child (mean = 78.23). Therefore, we only used IBM-Watson for diarisation.

2.3. Automatic word recognition with UNSW ASR system

Word recognition is conducted using the UNSW ASR engine [10]. The UNSW ASR engine's acoustic model is based on deep-learning and trained on ∼400 hours of children's speech from four different American speech corpora using the Kaldi toolkit [9, 10]. The UNSW ASR engine uses a large-vocabulary language model trained on transcriptions of adult and child speech to cope with the spontaneous conversations occurring in the recording, and prioritises the 130 target words of Task 1.

Audio files with IBM-Watson textgrids containing diarised and time-aligned transcription are fed to the UNSW ASR. The UNSW ASR returns word-level transcriptions for all intervals identified by IBM-Watson, replacing IBM-Watson's transcriptions with its own. Word-level boundaries are compared to the prompt interval (Sec. 2.1). Prompt intervals that do not overlap with any word intervals are processed by the UNSW ASR tool and the intervals of the recognised words are added to all speaker tiers. On the test set of the same four children, UNSW ASR achieved a word error rate of 18%-45% (mean = 30%), less than half of the word error rate of IBM-Watson.

To minimise manual annotation time, IBM-Watson is used

to isolate child only segments and the UNSW ASR tool to orthographically transcribe audio. The IBM-Watson web service and the UNSW ASR are linked with a Python script taking several sound files and returning diarised and transcribed textgrids.

3. Manual correction

Manual correction is required to improve the accuracy of the transcription and placement of target words on the automatically generated textgrids as well as to remove annotation for non-target words. Given the large size of the collected corpus, a team of annotators based at multiple sites is involved in the manual correction process.

3.1. Distributing data across multiple sites

A web tool was developed to distribute the audio recordings and the automatically generated textgrids across the annotation team, collect the manually corrected textgrids, and ensure consistency across annotators. The front-end of the web tool provides two interfaces: one for annotators and one for their supervisors. Users are directed to their corresponding interface based on their roles and authentication credentials.

The annotator interface provides options to download a new, automatically allocated audio file with the associated prompt and uncorrected speaker textgrids, as well as to upload the corrected annotation files. To test correction consistency, benchmark files are inserted at regular intervals. Benchmark files are files with ground truth annotations (i.e. previously corrected by expert phoneticians; Sec. 3.3). Annotators are blinded as to which files are benchmark files. When corrections to benchmark files are uploaded, the corrections are automatically scored against ground truth annotation (Sec. 3.3). The annotator cannot proceed to the next file unless a passing score is achieved. Feedback and additional training will be provided.

The backend of the web tool contains a database that keeps track of file allocation statistics, annotator scores, and account information. The backend logic is implemented in PHP scripting language with the MySQL database while the front end is developed with HTML5 and JavaScript. The web tool is hosted in a university web server which provides accessibility to the annotators from different locations.

3.2. Praat interface

A Praat [11] tool was designed to enable efficient manual correction of word-level annotation. The frontend streamlines the correction procedure with a user-friendly interface. The backend automates low-level and time-consuming tasks (e.g., opening and saving textgrids) and loads those portions of the sound file that are likely to contain a target word and skips the rest.

3.2.1. Steps of manual correction

Tasks that could not be fully automated are streamlined by creating a series of five tasks with simple instructions presented in Praat pop-up windows. As the IBM-Watson diarisation tool only identifies the speech segments belonging to the different speakers (interviewer, model speaker, child) and not who the actual speakers are (Sec. 2.2), the first manual task is identifying which tier belongs to the child. When the annotator starts a new file, the waveform and the spectrogram with a textgrid showing the prompts on one tier and the three speakers on three separate tiers are displayed. The annotator is asked to identify which tier is the best match for the child's speech. The annotator is instructed to scroll through the recording and to listen to as much of the audio as needed before making their decision.

After selecting the child's tier, the script proceeds to the correction phase of the selected tier, while the other two speaker tiers are discarded. The prompt tier is displayed together with the child's tier. Correction of an interval contains four steps: interval evaluation, label and boundary evaluation, noise evaluation, and phoneme-level discrepancy evaluation. When an interval is presented, the annotator has the options to Accept, Delete (not child), or Delete (not target) the interval. The annotator is instructed to only accept an interval if the interval contains the child's voice and the child's speech matches the prompt.

Once the annotator accepts an interval, the annotator is instructed to either accept or edit the interval's label (i.e., the automatic transcription) and/or its boundaries. Annotators are trained to transcribe the child's speech using standard English spelling and grammar, e.g., spell the target word *rhinoceros* with the letter "r", regardless of whether it was produced with the rhotic or an approximant, such as [w] or [ʋ]; and *two eggs* with the plural marker -s, even if the child produced *two egg*. Annotators are trained to ensure that the boundaries contain the entire word produced by the child and no other speech. In particular, they are warned to be careful to include final stop bursts which may be cut off by automatic annotation tools due to children producing a longer closure phase in stops than adults [12].

Once all the necessary corrections to the intervals are complete, the annotator is asked to decide whether the recording contains any noise (e.g., overlap between the interviewer and the child, background noise) and whether the token contains any phoneme-level insertions (e.g., *skirts* for *skirt*), deletions (e.g., *four egg* instead of *four eggs*), or substitutions (e.g., *tooth* produced with a final /f/ instead of /θ/). Annotators are told to flag any phoneme-level discrepancy from adult production irrespective of it being age-appropriate (e.g., [w] for /ɹ/ flagged at each instance, regardless of age). When a word is flagged as non-adult like, annotators must identify the differing phoneme(s). To prevent the annotators from spending too much time on noise and discrepancy evaluations, the script only allows the target word to be played twice for noise and twice for discrepancy evaluation (four times in total) and automatically closes the sound and the textgrid while waiting for the noise and discrepancy decisions. The annotator is able to move to the next interval once all four steps for a given interval are completed.

To improve annotation quality, the script checks that the annotator follows the instructions as closely as possible. A notification is triggered when the interval label does not match the prompt and/or when the annotator's evaluation does not match what they have done (e.g., the annotator has evaluated the automatically placed interval label as correct yet they have changed it). The annotator is instructed to correct their error(s). Erroneous edits are not saved and the annotator is not allowed to move onto the next interval until all checks are passed. When all checks are passed, final edits and progress are saved and the interval is marked as corrected, allowing the annotator to exit Praat any time. At the next launch, the script loads an unfinished sound file - textgrid pair at the first uncorrected interval.

Once the annotator corrects the textgrid for the entire sound file, a clean and corrected textgrid free from unnecessary annotations is created by removing all remaining child labels in which the label does not match the prompt. To create an easily searchable output, a csv file is generated that lists the label, and the start and end time of every on-prompt target and is saved automatically with the child's identifier. The clean and corrected textgrid, together with the csv file are uploaded to the web tool.

3.2.2. *Focus on intervals of interest*

The Praat tool automatically identifies those intervals that are the most likely to contain target words by comparing the prompt tier to the automatic transcription of the child's speech in an iterative process. Comparison of the child's production and the prompt is required as target words produced without being prompted are excluded from the data. For instance, frequent words (e.g., *yes, no, that*) produced without a prompt, or easily confusable words produced for the incorrect prompt (e.g., the target *boat* produced for the picture of a canoe) are excluded.

In the first iteration, the script loads the intervals in which the automatic transcription (Sec. 2.3) matches the prompt, as these are likely to contain a target word. For instance, the interval labelled with the target word *key* is loaded only if the prompt interval is also labelled with *key*. When more than one interval labels match the prompt, all are loaded and corrected. The tool tracks which prompts are found and corrected in the first iteration. All other intervals, i.e., intervals transcribed as non-target words or as target words not matching the prompt are skipped.

In the second iteration, the script loads all the word-level intervals that map onto prompts not found in the first iteration. For instance, if the target *key* was found in the first iteration, then all other word-level intervals that map onto the prompt *key* are assumed to be non-target and skipped. If, however, the target *key* was not found in the first iteration, all intervals identified as the child's speech produced during the prompt *key* are displayed, irrespective of their transcription. If *key* was produced twice, and transcribed as *he* and *e* respectively, both are loaded and corrected in the second iteration.

In the third iteration, the script loads an entire prompt interval if the prompt was not found in either of the previous iterations. The annotator is asked to listen to the entire interval during which a prompt was displayed to identify and add the target word. More than one target interval can be added. We assume that target words that are not found in any of the iterations were not produced or were only produced when not prompted.

Repeated targets are identified when they are corrected in the same iteration. Repetitions are skipped when they would be corrected in different iterations, e.g., if one instance of *key* has a correct automatic transcription and the other does not, the one with the correct automatic transcription is identified in the first iteration, and the other is skipped in the second iteration.

The length of audio the annotator must listen to is reduced. For instance, a prompt might be displayed for 6-10 seconds; however, the word-level interval(s) might be only 0.6-1 second long. The annotator only listens to the automatically generated, shorter word-level interval(s). In a sample of four children, the length of audio the annotator listened to was reduced from a total of 107 minutes to 22 minutes. We identified 125–127 targets out of a total of 130 (mean = 126). 75–99 targets (mean = 85.75) were found in the first iteration, 15–30 (mean = 23) in the second, and 12–23 (mean = 17.25) in the third.

3.3. Reliability checks

Interrater reliability is typically done by all annotators correcting the same set of files (customarily 20% of the data). Due to the large number of files (approximately 750 sound files for 750 children), a 20% cross-correction is not possible as it would have required 140 speakers being marked by all annotators. Similarly, 20% rescoring for intrarater reliability is not feasible due to the large number of files.

Therefore, a ground-truth approach is chosen to achieve consistency. Eight benchmark sound files are identified consisting of four older (10-12 years, M = 2, F = 2) and four younger children (3-5 years, M = 2, F = 2). Four of the children are typically developing, and four have current speaking disorders (one older male, one younger male, one older female, one younger female). Four expert phoneticians manually correct the benchmark files independently from each other and compare their correction. In case of a disagreement, consensus is reached through discussion; disagreement typically arises regarding flagging non-adult like productions and almost never regarding identifying target words.

In every benchmark file, the annotator's work is compared to the ground truth by comparing the number of targets identified and calculating overlap rate for matching targets between the benchmark and the current annotation. The number of identified targets tests whether the annotator found all the target words in the audio data. Repetitions of targets are not counted towards the pass rate, as the task was not designed to capture repetitions for targets and the Praat interface does not require identifying all repetitions. Overlap rate is calculated relative to the ground truth annotation, using the time shared between ground truth and current annotation (*Dur Shared*), the duration of the ground truth annotation (*Dur GT*), and the duration of the current annotation (*Dur Current*) (Equation 1), [13, 14].

$$Overlap = \frac{DurShared}{DurGT + DurCurrent - DurShared} \quad (1)$$

Overlap rate ranges from 1 (complete agreement, 100% overlap) to 0 (no agreement, 0% overlap) and penalises too long and too short intervals equally. If the ground truth annotation for a target is 0.6s long, a current annotation with 1.2s duration and 0.6s shared duration and a current annotation with 0.3s duration and 0.3 shared duration both yield an overlap rate of 0.5.

Passing rate for annotators is calculated automatically from the number of target words identified and from the overlap rate (Sec. 3.1). If a passing rate is not achieved, the corrected files after the last successful checkpoint (if any) will be reviewed (Bottom yellow arrow in Fig. 1) and re-corrected if needed.

4. Conclusions and future work

Our goal was to provide time-aligned orthographic transcription to the prompted single word elicitation task (Task 1) in the AusKidTalk corpus. We overcame the lack of suitable ASR tools required for the task by developing a semi-automated workflow that concatenates IBM-Watson diarisation with the task-specific UNSW ASR orthographic transcription system to automatically generate textgrids with time-aligned transcription. A webtool distributes the automatically generated textgrids and collects manually corrected textgrids, and implements consistency checks against ground truth annotations. Correction is done in a custom Praat interface.

This workflow is essential to achieve an orthographic annotation of target items in an efficient manner and creates a corpus to be further processed using forced alignment to generate phonemic annotations. However, a lot of valuable data are necessarily disregarded, such as incidental conversations between the interviewer and the child. Therefore, raw data files will be made available as a corpus for researchers who are interested in more than just the target items. We aim to extend the workflow to include non-word repetition by adding the pseudowords to the UNSW ASR system's dictionary, and use the workflow as a starting point for annotating the sentence repetition task using a modified Praat interface.

5. Acknowledgements

This project was supported by the Australian Research Council LE190100187 and FT180100462 grants, as well as the University of New South Wales, The University of Sydney, Western Sydney University, Macquarie University and The University of Melbourne. We would like to thank our participants without whom this project would not have been possible.

6. References

[1] http://www.auskidtalk.edu.au/

[2] Ahmed, B., Ballard, K. J., Burnham, D., Tharmakulasingam, S., Mehmood, H., Estival, D., Baker, E., Cox, F., Arciuli, J., Benders, T., Demuth, K., Kelly, B., Diskin-Holdaway, C., Shahin, M., Sethu, V., Epps, J., Lee, C. B. and Ambikairajah, E., "AusKidTalk: An Auditory-Visual Corpus of 3-to 12-year-old Australian Children's Speech", ISCA, 2021.

[3] Schiel, F., Draxler, C., Baumann, A., Ellbogen, T. and Steffen A. "The production of speech corpora" Version 2.5, 2012, http://www.bas.uni-muenchen.de/Forschung/BITS/TP1/Cookbook

[4] Russell, M. and D'Arcy, S. "Challenges for computer recognition of children's speech" , Workshop on Speech and Language Technology in Education, 2007.

[5] Elenius, D. and Blomberg, M. "Adaptation and normalization experiments in speech recognition for 4 to 8 year old children", Interspeech, 2749–2752, 2005.

[6] Keshet, J., "Automatic speech recognition: A primer for speech-language pathology researchers", International Journal of Speech-Language Pathology, 20(6):599–609, 2018.

[7] Chen, N. F., Tong, R., Wee, D., Lee, P. X., Ma, B. and Li, H., "SingaKids-Mandarin: Speech Corpus of Singaporean Children Speaking Mandarin Chinese", Interspeech, 1545-1549, 2016.

[8] Transcribing speech with Watson Speech to Text. (2021, 2022). IBM Corporation. https://www.ibm.com/docs/en/cloud-paks/cp-data/4.0?topic=solutions-watson-speech-text Accessed: 5 September 2021

[9] Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., Silovský, J., Stemmer, G. and Veselý, K. "The Kaldi speech recognition toolkit", IEEE workshop on automatic speech recognition and understanding, 2011.

[10] Shahin, M. A., Lu, R., Epps, J. and Ahmed, B. "UNSW System Description for the Shared Task on Automatic Speech Recognition for Non-Native Children's Speech", Interspeech, 265-268, 2020.

[11] Boersma, P. and Weenink, D. (2022). Praat: doing phonetics by computer [Computer program]. Version 6.2.14, retrieved 24 May 2022 from http://www.praat.org/

[12] Millasseau, J., Ivan, Y., Bruggeman, L. and Demuth, K., "Acoustic cues to coda stop voicing contrasts in Australian English-speaking children", Journal of Child Language, 48(6): 1262-1280, 2021.

[13] González, S., Grama, J. and Travis, C., "Comparing the accuracy of forced-aligners for sociolinguistic research", Linguistics Vanguard, 6(1).

[14] Paulo, S. and Oliveira, L. C., "Automatic phonetic alignment and its confidence measures", Proc. Advances in Natural Language Processing, 36–44, 2004.

Multi-Task Learning for Speech Attribute Detection of Children's Speech

Mostafa Shahin, Beena Ahmed, Julien Epps

School of Electrical Engineering and Telecommunications, University of New South Wales, Sydney, Australia

mostafa_shahin@ieee.org, beena.ahmed@unsw.edu.au, j.epps@unsw.edu.au

Abstract

Speech attributes, including manners and places of articulation, provide a detailed description of sound production directly related to the speech articulators. An accurate modeling of such attributes enriches automatic pronunciation assessment applications by providing informative feedback to the user. In this paper we propose a DNN based classification model structure to automatically detect the absence or existence of 25 specific attributes in English children's speech. In the models, we utilised multi-task learning (MTL) with frame-level phoneme classification as an auxiliary task and a discriminative additive (DA) task for highly confusable phonemes to improve the detection of attributes. We compared the performance of the 25 DNN-MTL and DNN-MTL-DA attribute detection models across all the phonemes to base DNN models in two different children corpora to determine the impact of MTL and the DA task. We also compared the performance of our DNN-MTL-DA across different children's age groups. Our attribute models achieved detection accuracies ranging from ~80% to ~91%, with the best detection accuracy for *Nasal* and *retroflex* and worst for *Tense* and *affricates*.

Index Terms: speech attributes, children speech, multi-task learning

1. Introduction

Unlike acoustic modeling for automatic speech recognition (ASR) applications where an abstract model that can handle all variations of the same word is desirable, pronunciation assessment applications such as L2 learning, speech therapy and language proficiency tests, require the accurate detection of any deviation from standard pronunciation.

Speech attributes, such as manners and places of articulations, provide a low-level description of sound production in terms of which articulators are involved and how these articulators move to produce a specific sound. Any alteration in these attributes causes a pronunciation error. Therefore, an accurate modeling of these attributes can pave the way to a fully automated and interactive pronunciation assessment application where the learner receives formative and diagnostic automatic feedback not only about the existence of incorrect pronunciation but also how the error is made. Furthermore, modeling speech attributes can be performed solely using standard pronunciation datasets which are abundantly available compared to non-standard datasets. Additionally, speech attributes are common among most spoken languages enabling modelling with speech corpora from multiple languages.

Speech attributes have been successfully utilized in various domains. In the context of ASR, Lee et al. proposed a bottom-up speech recognition approach based on a bank of speech attribute detectors [1]. In [2] authors used speech attribute-based model to rescore a word-lattice generated by Gaussian mixture model (GMM) and deep neural network (DNN) based ASR system. [3] introduced an i-vector representation of manners and places of articulation attributes for the detection of foreign accent speech in Finnish and English languages. The system achieved a 15% error reduction when compared to a spectrum-based technique.

In the pronunciation assessment domain, [4] used scores derived from speech attribute models to assess the pronunciation quality of L2 Mandarin learners. The system outperforms phoneme-based goodness of pronunciation technique (GOP) by reducing the equal error rate by ~9% relative. In [5], authors proposed an anomaly detection-based pronunciation verification system by training phoneme-specific one-class support-vector machine (OCSVM) model using speech attribute features. However, most previous work has focused on speech attributes in adult speech. Children's speech poses an additional challenge due to its high inter- and intra-speaker variability compared to adult speech.

In this paper we propose a deep learning approach to accurately detect 25 speech attributes including manners and places of articulations in English children's speech. In this approach we used 25 separate binary output DNN-based models to detect the absence or existence of each attribute trained two public speech corpora including speech from children aged 5 to 15 years old. In the models, we used phoneme classification as a secondary task in a multi-task structure (DNN-MTL) and additional discriminative tasks (DNN-MTL-DA) between confusable phonemes to improve the performance of the speech attribute detectors. We conducted experiments to validate the performance of the DNN-MTL models in detecting all 25 attributes across all phonemes and over different age groups.

The rest of the paper is organized as follow. The method and the speech corpora used are explained in Section 2. Section 3 represents the experiments and the analysis of the results. Finally, conclusions are drawn in Section 4.

2. Method

2.1. Speech Corpora

Two publicly available speech corpora were utilized in this work, the Oregon Graduate Institute (OGI) kids' speech corpus [6] and Colorado University (CU) Kid's prompted, read and summarized speech corpus [7, 8].

Table 1. *The distribution of the speech corpora.*

	OGI			CU		
	Train	**Valid**	**Test**	**Train**	**Valid**	**Test**
Speakers	794	162	162	716	107	94
Segments	~42k	~8.5k	~8.5k	~66k	~9k	~8.3k
Hours	59.4	7.7	7.6	59	7.3	7.6

The OGI corpus contains 11 age groups from kindergarten to grade 10 while the CU corpus consists of 6 age groups from kindergarten to grade 5. Each dataset was split into three subsets for training, validation and testing as demonstrated in Table 1. The data was split to ensure that each subset included all age ranges.

In our experiments we used each speech corpus separately to show the robustness of speech attribute detection over different domains. The OGI corpus includes recordings from across a wider age range while the CU corpus has a larger vocabulary and more spontaneous speech.

2.2. Speech Attributes Detectors

We adopted 25 speech attributes representing mainly the manners and places of articulations in addition to other phonetic characteristics such as voiceless, roundness, tenseness, etc. Table 2 shows the list of speech attributes along with their associated phonemes in ARPAbet format [9].

Table 2. *List of speech attributes.*

	Attribute	Phonemes
Manners	Vowel	iy ih eh ey ae aa aw ah ao oy ow uh uw er
	Semivowel	y w
	Fricative	jh ch s sh z zh f th v dh hh
	Nasal	m n ng
	Stop	b d g p t k
	Approximant	w y l r
	Affricates	ch jh
Places	Coronal	d l n s t z
	High	ch ih iy jh sh uh uw y ow g k ng
	Dental	dh th
	Glottal	hh
	Labial	b f m p v w
	Low	aa ae aw ay oy
	Mid	ah eh ey ow
	Retroflex	er r
	Velar	g k ng
Others	Anterior	b d dh f l m n p s t th v z w
	Back	ay aa ah ao aw ow oy uh uw g k
	Continuant	aa ae ah ao aw ay dh eh er r ey l f ih iy oy ow s sh th uh uw v w y z
	Round	aw ow uw ao uh v y oy r w
	Tense	aa ae ao aw ay ey iy ow oy uw ch s sh f th p t k hh
	Voiced	aa ae ah aw ay ao b d dh eh er ey g ih iy jh l m n ng ow oy r uh uw v w y z
	Liquids	l r
	Monophthong	ao aa iy uw eh ih uh ah ae
	Diphthong	ey ay ow aw oy

For each attribute, a binary fully connected DNN model was trained to classify each frame as +ve or -ve when the underlying attribute was detected or not respectively. We further used a frame-level phoneme classification auxiliary task to improve the generalization of the main attribute detection task in a MTL scenario. In some specific attributes, additional task/s were added to force the network to discriminate between pairs of highly confused phonemes better that were described by attributes in opposite classes (+ve/-ve). The confused phonemes were determined from a phoneme confusion matrix computed using a frame-level phoneme classification model. The MTL model architecture is depicted in Figure 1.

The cross-entropy function was used to compute the loss in the attribute output as well as outputs of all other tasks. The total network loss was computed as follows:

$$L_A(\theta) = -\sum_i \log P(y_A^{(i)}/x^{(i)};\theta) \quad (1)$$

$$L_P(\theta) = -\sum_i \log P(y_P^{(i)}/x^{(i)};\theta) \quad (2)$$

$$L_D(\theta) = -\sum_i log\, P(y_D^{(i)}/x^{(i)};\theta) \quad (3)$$

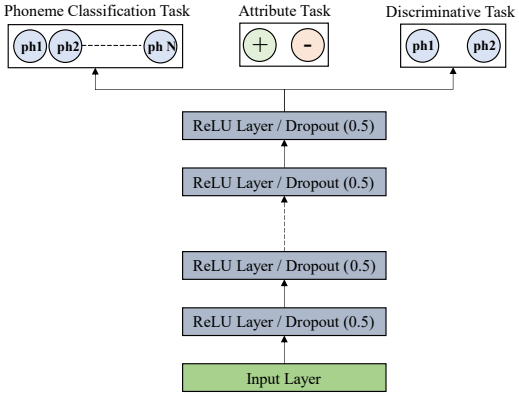


Figure 1: *Multi-Task learning model architecture.*

Where L_A, L_P and L_D are the losses computed from the attribute, phoneme classification and discriminative tasks respectively, while $y_A^{(i)}, y_P^{(i)}$ and $y_D^{(i)}$ are labels of sample $x^{(i)}$ for the three outputs. The total loss is a weighted sum of the three losses as follow:

$$L(\theta) = L_A(\theta) + \lambda_P L_P(\theta) + \lambda_D L_D(\theta) \quad (4)$$

Where λ_P and λ_D are the weights of the losses at the phoneme classification and the discriminative tasks, respectively.

From each frame, 26 filter banks were extracted along with their delta and acceleration components. N consecutive frames were spliced together to form the final feature vector fed to the DNN model. The number of layers, the number of units per layer, initial learning rate and the N spliced frames were empirically determined. The rectified linear unit (ReLU) was used as an activation for the hidden units while the softmax used for the output units of all tasks. The training ran for a maximum of 50 epochs with early stopping by monitoring the validation loss of the attribute detection task. Moreover, the dropout regularization of 0.5 was used to alleviate the overfitting effect and the update of the parameter was optimized using the Adam technique.

In the training of each speech attribute model, frames from phonemes belonging to the underlying attribute were labeled as +ve while all other phonemes were labeled as -ve. To avoid the model being biased, we balanced the data by selecting an equal number of samples in each class. We ensured that samples were selected from all phonemes of +ve and -ve groups. The time boundaries of the phonemes were obtained by forced alignment using a children's ASR system [10].

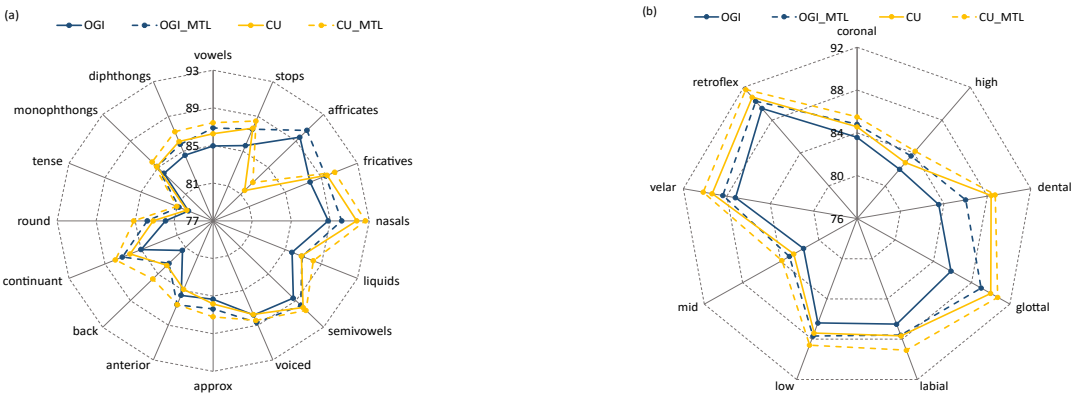


Figure 2: *Detection accuracy of (a) manners of articulations and other phonetic attributes and (b) places of articulation attributes for OGI and CU datasets.*

3. Results

The baseline DNN model consists of one output layer for attribute detection, and 5 hidden layers with 1024 ReLU hidden units each. Features from 5 frames were spliced to form the input feature vector. This baseline architecture is similar to the one used in [11] for adult speech attribute detections of the Wall Street Journal (WSJ) speech corpus [12]. Table 3 shows the classification accuracy of the baseline DNN models when trained and tested against OGI speech corpora and CU speech corpora. The last column shows the classification accuracy of a similar DNN model applied to the adult WSJ speech corpus obtained from [11].

Table 3. *Classification accuracy of different speech attribute detectors trained and tested using children's speech corpora (CU and OGI) and adult speech corpus (WSJ).*

Attribute	OGI	CU	WSJ [11]
anterior	85.6	84.9	92.5
approximant	85.3	85.8	96.4
back	81.5	83.7	93.1
continuant	85.0	86.2	93.5
coronal	83.5	84.6	92.4
dental	83.6	88.4	99.0
fricative	87.7	89.6	96.2
glottal	85.8	90.0	99.7
high	82.0	82.8	95.0
labial	86.5	87.7	96.9
low	86.4	87.3	96.9
mid	81.6	82.6	93.8
nasal	88.8	91.7	97.7
retroflex	89.4	90.8	98.5
round	81.9	83.1	94.9
stop	85.6	87.6	95.7
tense	80.7	81.0	90.6
velar	87.2	89.3	98.7
voiced	87.8	87.7	95.3
vowel	85.0	86.2	92.8
Average	85±2.5	86.6±2.9	95.5±2.5

A separate model was trained for the detection of each attribute. Approximately 15 hours of speech from the WSJ corpus was used to training the adult speech attribute detection models [11]. Despite training the CU and OGI speech attribute detection models with almost 4 times as much training data, the classification accuracies of both children's datasets over all the speech attributes were much lower than the corresponding accuracies achieved using models for the adult WSJ dataset as shown in Table 3. The *tense* attribute was shown to be the most difficult attribute to be automatically detected for both adult and children. The most accurate attributes in adult are *glottal* followed by *dental* and *velar* while in children, *nasal* achieved the highest detection accuracy followed by *retroflex* and *fricative*. These results demonstrate the increased difficulties of detecting speech attributes in children's speech compared to adult speech.

Figure 2 depicts the results on both OGI and CU datasets using the baseline DNN model and the proposed DNN-MTL model. The best parameters of the MTL model were 5 hidden layers with 2048 hidden units each, 0.001 initial learning rate, and 5 spliced windows.

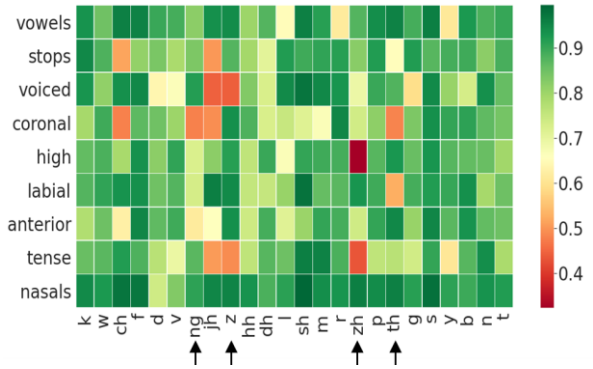


Figure 3: *Attribute detection accuracy over consonant phonemes for CU dataset. /zh/ has very low accuracies in tense and high attributes and high accuracies in nasals and labial. Similar behavior for /ng/ in coronal compared to voiced, and /th/ in labial and tense compared to vowel.*

Overall, results with the CU dataset are better than with OGI. Although both models were trained with almost the same amount of data, the CU dataset has more variations in vocabulary and more continuous speech compared to OGI. As shown in the figure, the detection accuracy ranges from 80% to almost 91% for both datasets. *Nasal*, *retroflex* and *semivowel* attributes achieved the highest detection accuracy of around 91% followed by *glottal*, *fricatives* and *velar* with nearly 90% correct detection rate. *Tense* has the lowest detection accuracy of ~80% followed by *affricates*, mid and high of ~82%.

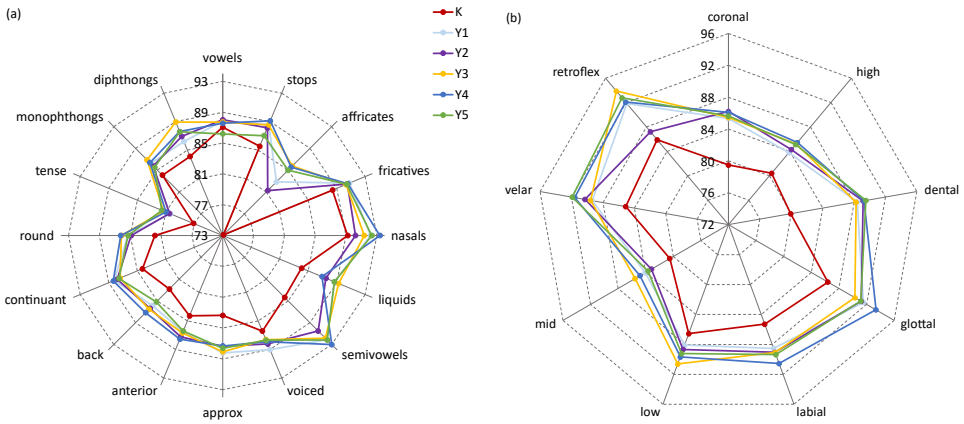


Figure 4: *Distribution of detection accuracy of (a) manners of articulations and other phonetic attributes and (b) places of articulation attributes of CU dataset over different age group from Kindergarten(K) to year 5 (Y5).*

It is obvious from the figure that using the phoneme classification auxiliary task improved the performance of all speech attributes in both CU and OGI datasets with an absolute increase in the accuracy between 0.5% to as high as ~3%. The detection error of the OGI glottal attribute decreased from 14% to around 11% while the CU back attribute detection error decreased from ~16% to ~14%.

We further broke down the results of each attribute by all +ve and -ve phonemes for the CU dataset. Figure 3 shows a heatmap of a subset of this breakdown focusing on the consonants. It is noticeable from the figure that the performance is inconsistence over different phonemes. For instance, /zh/ suffers from low classification accuracy of ~40% and ~35% in *tense* and *high* attributes respectively while *nasals* and *labial* obtained ~97% and ~93% respectively. To better understand this behaviour, we computed the frame-level phoneme confusion matrix using a phoneme classification DNN model trained on the same CU dataset.

Table 4 lists the most confused phonemes for consonants that obtained low attribute classification accuracy as indicated in Figure 3. We concluded that when two highly confused phonemes are in opposite sides of specific attributes (i.e one +ve phoneme and the other one -ve phoneme), the accuracy of one or both degraded significantly. For example, /ng/ has a high confusion with /n/, in *labial* both /n/ and /ng/ are positive samples and hence they achieved a high accuracy > 90%. On the other hand, in *coronal* and *anterior*, where /ng/ is -ve and /n/ is +ve, the accuracy of /ng/ degraded to ~49% and ~62%, respectively. The same behaviour was present with /th/ and /f/ in *labial* and *coronal* and /z/ and /s/ in *tense* and *voiced*.

Table 4. *Phoneme confusion*

Phoneme	Confused phonemes (misclassification rate)
/zh/	/sh/ (37%), /jh/ (7%)
/z/	/s/ (38%), /t/ (2%)
/ch/	/sh/ (23%), /t/ (22%)
/ng/	/n/ (34%), m (4%)
/jh/	/t/ (17%), ch (10%), sh (10%)
/th/	/s/ (19%), f (15%), t (7%)

To improve the performance of the model in these specific phonemes, we investigated the use of a discriminative task in the learning of the attribute detection model to guide the model, so it discriminates better between highly confused phonemes. In the training of the anterior detector, we added an additional output to discriminate between /ng/ and /n/ achieving ~8% absolute increase in the *anterior* attribute detection of /ng/ frames from ~62% to ~70%. The same architecture was used in training the *labial* detector; adding a /th/ and /f/ discrimination output, also improved the /th/ accuracy from ~66% to ~71%.

Figure 4 represents a further analysis of the results over 6 age groups from the CU dataset from Kindergarten (K) to Year 5 (Y5). Except for *vowels*, Kindergarten children show a low detection accuracy for almost all other attributes with the highest degradation notice in *affricates*. This can be explained by the rapid change in the articulators between younger and older children and the development of speech sounds. Generally, *vowels* are among the first phonemes to be mastered by children at very young ages. On the contrary, a pronunciation error of the *affricate* sounds known as de-affrication is common in children till 5 years old. This is where the child tends to replace affricates like /ch/ and /jh/ with fricative or stop like /sh/ or /d/.

4. Conclusions

In this work, 25 binary DNN-based models were trained to detect different speech attributes, including manners and places of articulation, in children's English speech. Two publicly available datasets were employed, namely CU and OGI corpora, for the training and evaluation of the models.

The speech attribute models achieved detection accuracies ranging from ~80% up to 91% where *nasal*, *retroflex* and semivowels obtained the highest accuracy. A frame-level phoneme classification auxiliary task was further explored to improve the generalization of the speech attribute models attaining ~20% average reduction in the detection error.

A phoneme-level breakdown of the detection error shows a significant increase in the error rate in highly confused phonemes when belonging to different classes of specific attribute. An addition discriminative task was added to the model architecture to alleviate this effect by discriminating between confused phonemes. When applied to /th/ and /f/ in the detection of *labial* attribute and to /ng/ and /n/ in the detection of anterior, the method achieved a 21% and 14% reduction in the classification error of /ng/ and /th/ respectively.

A further age group performance analysis was conducted demonstrating an obvious degradation in the detection accuracy at the young age group (Kindergarten) for most of the speech attributes except for *vowels* where the performance was consistent amongst all age groups.

5. References

[1] C.-H. Lee, and S. M. Siniscalchi, “An information-extraction approach to speech processing: Analysis, detection, verification, and recognition,” *Proceedings of the IEEE,* vol. 101, no. 5, pp. 1089-1115, 2013.

[2] I.-F. Chen, S. M. Siniscalchi, and C.-H. Lee, "Attribute based lattice rescoring in spontaneous speech recognition." pp. 3325-3329.

[3] H. Behravan, V. Hautamäki, S. M. Siniscalchi, T. Kinnunen, and C. Lee, “i-Vector Modeling of Speech Attributes for Automatic Foreign Accent Recognition,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing,* vol. 24, no. 1, pp. 29-41, 2016.

[4] W. Li, S. M. Siniscalchi, N. F. Chen, and C. Lee, "Improving non-native mispronunciation detection and enriching diagnostic feedback with DNN-based speech attribute modeling." pp. 6135-6139.

[5] M. Shahin, and B. Ahmed, “Anomaly detection based pronunciation verification approach using speech attribute features,” *Speech Communication,* vol. 111, pp. 29-43, 2019.

[6] K. Shobaki, J.-P. Hosom, and R. A. Cole, "The OGI kids' speech corpus and recognizers."

[7] R. Cole, P. Hosom, and B. Pellom, *University of colorado prompted and read childrens speech corpus*, Center for Spoken Language Research, University of Colorado, Boulder, 2006.

[8] R. Cole, and B. Pellom, *University of colorado read and summarized story corpus*, Center for Spoken Language Research, University of Colorado, Boulder, 2006.

[9] A. Klautau. "ARPABET and the TIMIT alphabet," https://web.archive.org/web/20160603180727/http://www.laps.ufpa.br/aldebaro/papers/ak_arpabet01.pdf.

[10] M. Shahin, R. Lu, J. Epps, and B. Ahmed, "UNSW System Description for the Shared Task on Automatic Speech Recognition for Non-Native Children’s Speech."

[11] D. Yu, S. M. Siniscalchi, L. Deng, and C.-H. Lee, "Boosting attribute and phone estimation accuracies with deep neural networks for detection-based speech recognition." pp. 4169-4172.

[12] D. B. Paul, and J. Baker, "The design for the Wall Street Journal-based CSR corpus."

Linear Transformation from Full-Band to Sub-Band Cepstrum

Frantz Clermont

School of Culture, History and Language, Australian National University, Canberra, Australia
& Forensic Speech and Acoustics Laboratory, J.P. French Associates, York, England
dr.fclermont@gmail.com

Abstract

This paper demonstrates the possibility of estimating the cepstrum for a sub-band region of the full-band spectrum by a linear transformation of the full-band cepstrum. The parametric formulation of the transformation allows the selection of any sub-band within the full-band's frequency range. The result of the transformation is a Fourier series of band-limited cepstral coefficients (BLCCs) representing the selected sub-band. In practice, the upper bound of the BLCC series may be fixed at $(M \times W)$ without significant loss in spectral resolution, where M is the number of full-band coefficients, and W is the fraction of the full-band spectrum occupied by the sub-band's width.
Index Terms: cepstrum, full band, sub-band, Fourier series, linear transformation

1. Introduction

Sub-band cepstra are commonly estimated using the filter-bank method [1,2]. The bandpass filters are designed with specific widths, and their locations along the frequency axis are also pre-determined. While these constraints have not impeded progress in speech and speaker classification, they restrict the ability to explore the full potential of sub-bands.

Here we propose an alternative method, which bypasses the problem of re-adjusting the filter-bank design and obviates the cost of analysing the speech signal for every sub-band of interest. Our method requires only the full-band cepstrum and a linear transformation to contain it within the limits of any sub-band.

Sections 2 and 3 develop the formulae to transform full-band into band-limited cepstral coefficients (BLCCs). Section 4 illustrates these for two sub-bands selected from speech cepstra. The key features of our band-limiting method are summarised in Section 5 with a brief overview of application possibilities.

2. Full-band cepstrum

The log-magnitude spectrum $S(\omega)$ can be defined as a Fourier cosine series of the so-called cepstral coefficients C_k:

$$S(\omega) = \sum_{k=1}^{M} C_k \cos(k\omega), \quad 0 \le \omega \le \pi \tag{1}$$

Truncating this series after M terms yields a cepstrally-smoothed representation of $S(\omega)$ across the entire available bandwidth. The C_k are thus interpreted as full-band coefficients.

3. Band-limited cepstrum (BLC)

3.1. Band-limiting parameters

Let ω_1 and ω_2 be the limits of a sub-band, and let the associated change of variables $\omega \rightarrow \omega'$ be as follows:

$$\omega' = \pi \left[\frac{(\omega-\omega_1)}{(\omega_2-\omega_1)}\right], \quad \omega_1 \le \omega \le \omega_2 \tag{2}$$

From Eq. (2) it is easy to express ω as a function of ω':

$$\omega = \omega_1 + \left[\frac{(\omega_2-\omega_1)}{\pi}\right]\omega' = \omega_1 + W\omega' \tag{3}$$

where $0 \le \omega' \le \pi$, and where the scalar W is the ratio of the sub-band's width to the full-band's frequency range:

$$W = \left[\frac{(\omega_2-\omega_1)}{\pi}\right], \quad 0 < \mathrm{W} \le 1 \tag{4}$$

3.2. BLC formulation

The band-limited analogue of Eq. (1) is defined in Eq. (5) as a Fourier cosine series representing the spectral region of the full band delimited by the sub-band:

$$S\big(\omega(\omega')\big) = C'_0 + \sum_{l=1}^{N} C'_l \cos(l\omega'), \quad 0 \le \omega' \le \pi \tag{5}$$

The notation $\omega(\omega')$ indicates that the argument ω is a band-dependent function of ω' via Eq. (3), C'_l is the l-th band-limited cepstral coefficient (BLCC), and N is the upper bound of the BLCC series. The zeroth-order coefficient retains the level of the band-limited spectral region, and the other coefficients account for the spectral shape. While N theoretically goes to ∞, a much lower bound is sufficient to preserve the overall shape. This is discussed in Section 3.4 and illustrated in Section 4.

3.3. BLCC derivation

3.3.1. Fourier cosine coefficients: Formulae & solutions

The cepstral coefficients C'_l in Eq. (5) are derived from the standard Fourier integration formulae, as follows:

$$C'_{l=0} = \frac{1}{\pi}\int_0^{\pi} S\big(\omega(\omega')\big)\, d\omega' \tag{6}$$

$$C'_{l>0} = \frac{2}{\pi}\int_0^{\pi} S\big(\omega(\omega')\big)\cos(l\omega')\, d\omega' \tag{7}$$

$S\big(\omega(\omega')\big)$ is next replaced with the cosine series in Eq. (1) and ω with the right-hand side of Eq. (3) to yield:

$$C'_{l=0} = \frac{1}{\pi}\int_0^{\pi}\{\sum_{k=1}^{M} C_k \cos[k(\omega_1 + W\omega')]\}\, d\omega' \tag{8}$$

$$C'_{l>0} = \frac{2}{\pi}\int_0^{\pi}\{\sum_{k=1}^{M} C_k \cos[k(\omega_1 + W\omega')]\}\cos(l\omega')\, d\omega' \tag{9}$$

The finite sum over cepstral coefficients in Eqs (8) and (9) justifies reversing the order of integration and summation. The former becomes a separate operation, the result of which is inserted into the sum as a weighting coefficient. This is reflected in Eq. (10), where the band-limited coefficients C'_l are expressed as a weighted sum of the full-band coefficients C_k:

$$C'_l = \sum_{k=1}^{M} a_{lk} \cdot C_k, \quad l = 0,1,\ldots,N \tag{10}$$

The formulae for a_{lk} are expressed below as functions of ω_1 and ω_2. There are 3 cases depending on certain values of l:

if $l > 0$ and $l \neq kW$,

$$a_{lk} = \frac{2(kW)}{\pi[l^2-(kW)^2]}[(-1)^{l+1}\sin(k\omega_2) + \sin(k\omega_1)] \tag{11a}$$

if $l > 0$ and $l = kW$,

$$a_{lk} = \cos(k\omega_1) \tag{11b}$$

if $l = 0$,

$$a_{lk} = \frac{1}{k(\omega_2-\omega_1)}[\sin(k\omega_2) - \sin(k\omega_1)] \tag{11c}$$

Equation (11a) follows from evaluating the integral in Eq. (9) with adaptations of trigonometric solutions given in [3]. Equation (11b) is simply the limit of Eq. (11a) as $l \to$ kW. Equation (11c) also flows from Eq. (11a) by substituting l for 0 and dividing the result by 2. To decide which formula to use for $l > 0$, it is numerically desirable to test the condition $|(l - kW)| < \varepsilon$, where ε is a small positive number.

3.3.2. The weighted sum in matrix form: $\mathbf{c}' = \mathbf{Ac}$

The weighted sum in Eq. (10) implies a *linear transformation* from C_k to C'_l. Let $\mathbf{c}$ be a column vector $(M \times 1)$ of C_k, $\mathbf{c}'$ a column vector $((N+1) \times 1)$ of C'_l, and $\mathbf{A}$ the transformation matrix $((N+1) \times M)$ with elements a_{lk}. Equation (10) can thus be recast in the matrix form $\mathbf{c}' = \mathbf{Ac}$ laid out below. This is not only a convenient way of encapsulating Eqs (10) and (11), but one that is also useful for computer implementation.

$$\begin{bmatrix} C'_0 \\ C'_1 \\ \vdots \\ C'_l \\ \vdots \\ C'_N \end{bmatrix} = \begin{bmatrix} a_{0,1} & \cdots & a_{0,k} & \cdots & a_{0,M} \\ a_{1,1} & \cdots & a_{1,k} & \cdots & a_{1,M} \\ \vdots & & \vdots & & \vdots \\ a_{l,1} & \cdots & a_{l,k} & \cdots & a_{l,M} \\ \vdots & & \vdots & & \vdots \\ a_{N,1} & \cdots & a_{N,k} & \cdots & a_{N,M} \end{bmatrix} \begin{bmatrix} C_1 \\ \vdots \\ C_k \\ \vdots \\ C_M \end{bmatrix} \tag{12}$$

3.4. A practical bound for truncating BLCC series

Recall that Eq. (5) is a Fourier series representation of $S(\omega)$ over the interval $[\omega_1, \omega_2]$, just as Eq. (1) is one for the same function but over $[0, \pi]$. As a finite sum of sinusoids, $S(\omega)$ and its derivatives are continuous over $[0, \pi]$ and hence also over $[\omega_1, \omega_2]$. This guarantees that the series in Eq. (5) will converge uniformly for increasing values of the upper bound N. A relevant question is how large N needs to be in practice.

Our reasoning for truncating the BLCC series is as follows. If M cepstral coefficients represent the full-band spectrum with a certain resolution, then roughly the same resolution for the sub-band region should be achievable with $N = (M \times W)$, hereafter referred to as MW. This means retaining from M the fraction W of the full band occupied by the sub-band's width. Note that MW will generally not be an integer, and hence it must be rounded to be useable as a coefficient index.

Section 4.1 illustrates the BLCC series for two selected sub-bands, one narrower than the other. Section 4.2 shows the corresponding spectral fits around $N = MW$. In Section 4.3, a formula for estimating the truncation error is derived and applied to the same sub-bands.

4. Numerical illustrations

4.1. A glimpse at BLCC series for two sub-bands

Table 1 lists three sets of cepstral coefficients which were extracted at the frame marked vertically in Fig. 1. The full-band C_k (order M=14) were obtained by discrete-cosine transform of the log-magnitude FFT spectrum ranging from 0 to 5 kHz.

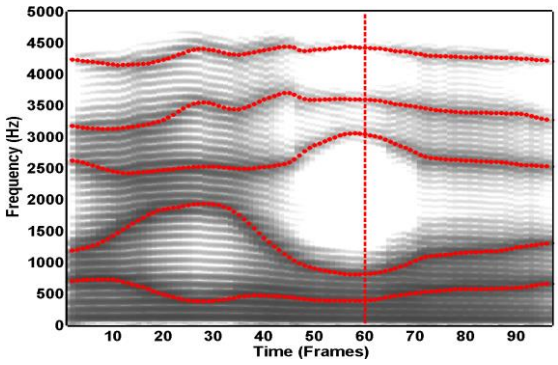


Figure 1: *Spectrogram of "IOWA" with superimposed formant-frequency tracks. A vertical line is marked at the frame of interest in this section and in Section 4.2.*

Equations (11) and (12) were used to generate BLCCs for these sub-bands: [0.1,0.9]-kHz and [2.5,4.0]-kHz. For the sake of illustration, the upper bound N of the BLCC series was set to 14, thus matching the number M of full-band coefficients and extending the BLCC series well beyond $MW = 2$ for the narrower sub-band and $MW = 4$ for the wider one.

Table 1. *Cepstral coefficients obtained at the frame marked in Fig. 1. The BLCCs C'_l for the two selected sub-bands are highlighted up to their respective MW.*

Coeff. index	Full-band C_k [0.0,5.0] kHz	BLCCs C'_l [0.1,0.9] kHz	BLCCs C'_l [2.5,4.0] kHz
0		2.5281	0.0865
1	0.9136	-0.1700	-0.2540
2	0.9127	-0.5929	-0.6726
3	1.3018	0.0108	-0.0979
4	-0.1548	-0.0898	-0.1701
5	-0.3611	0.0136	-0.0119
6	-0.2508	-0.0385	-0.0509
7	-0.2258	0.0077	-0.0087
8	-0.4771	-0.0214	-0.0253
9	0.0030	0.0048	-0.0054
10	-0.0895	-0.0136	-0.0155
11	0.1266	0.0032	-0.0036
12	-0.1663	-0.0094	-0.0106
13	-0.0302	0.0023	-0.0026
14	-0.0555	-0.0069	-0.0077

The zeroth-order C'_0 is much larger in the [0.1,0.9]-kHz range, which indicates a prominent region in the lower part of the spectrum. The next C'_l exhibit a consistent trend for both sub-bands: A drop in magnitude is noticeable after MW, with a subsequent decay of the BLCC series towards zero. This occurs 2 coefficients later for the sub-band [2.5,4.0]-kHz, whose width is about twice that of the narrower sub-band.

4.2. Fitting sub-band spectra with BLCCs

Is the proposed truncation after $N = MW$ detrimental to the spectral resolution in a sub-band region? To gain some insights into this question, full-band (in blue) and band-limited (in red), cepstrally-smoothed spectra are overlaid in Figs 2 and 3. These are based on the same C_k and C_l' listed in Table 1.

There is a major improvement in the spectral fit as N increases from 1 to 2 for both sub-bands. The fit then becomes very tight when $N = MW$, except at the edges where the zero slopes are likely to cause slow convergence of the BLCC series beyond MW. This may be inconsequential in practice, especially since only minor improvements are observed after MW.

In short, it could be said that the post-MW BLCCs contribute relatively little to the band-limited spectral shape. This is consistent with their decaying magnitudes noticed in Table 1.

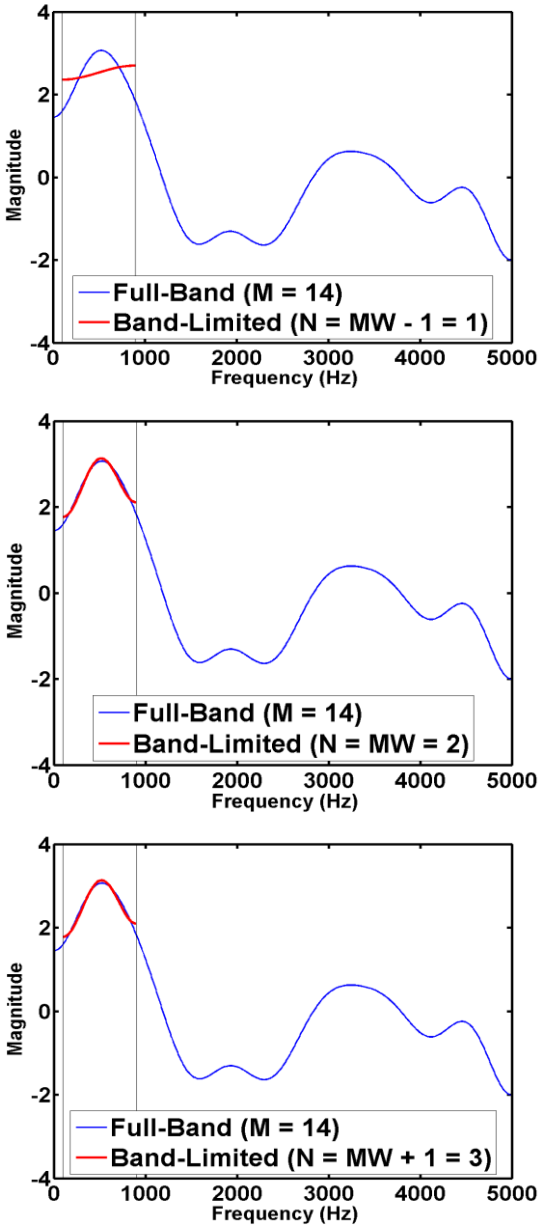


Figure 2: *Full-band & sub-band spectra based on C_k: [0.0,5.0]-kHz and C_l': [0.1,0.9]-kHz from Table 1.*

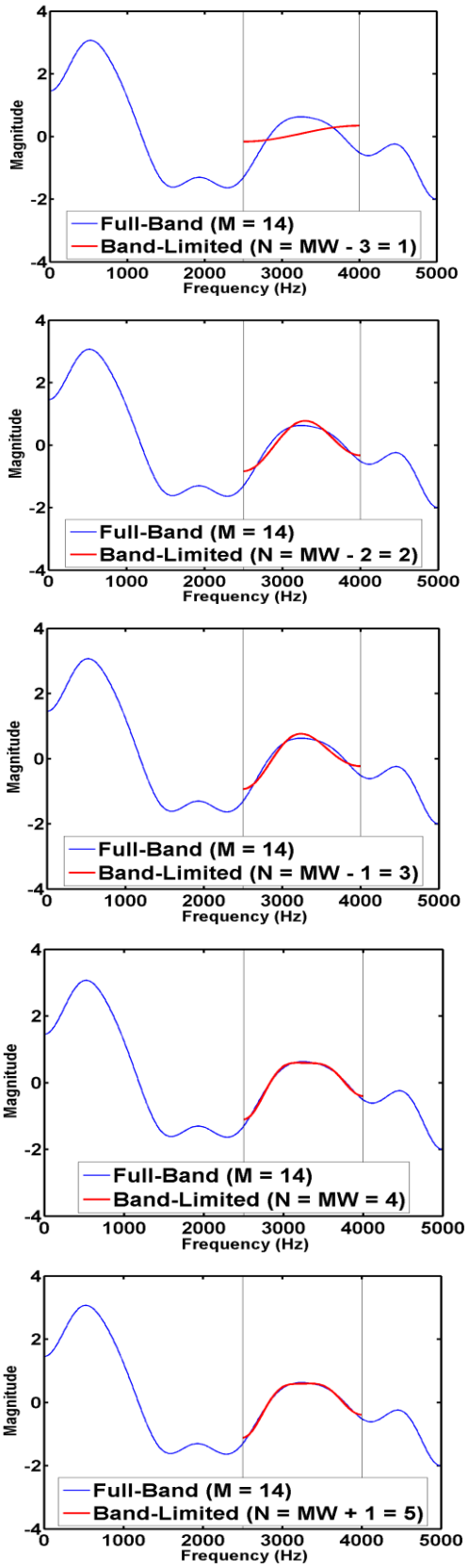


Figure 3: *Full-band & sub-band spectra based on C_k: [0.0,5.0]-kHz and C_l': [2.5,4.0]-kHz from Table 1.*

4.3. Truncation error estimation

The mean square of the Fourier cosine expansion of the BLCC series in Eq. (5) is exploited below to observe the truncation error before and after MW. The mean square formula for BLCCs is derived in Section 4.3.1 and an estimate of the error based on this formula is proposed in Section 4.3.2. In Section 4.3.3, the latter is applied to the "IOWA" example data.

4.3.1. Mean square over a sub-band

The mean square of $S(\omega(\omega'))$ is defined as:

$$\bar{S}^2 = \frac{1}{(\omega_2-\omega_1)}\int_{\omega_1}^{\omega_2}[S(\omega(\omega'))\,]^2 d\omega \qquad (13)$$

Recall Eq. (3) to express $d\omega$ as a function of $d\omega'$:

$$\omega = \omega_1 + W\omega', \quad 0 \le \omega' \le \pi,\ \omega_1 \le \omega \le \omega_2 \qquad (14a)$$

$$d\omega = W d\omega' = \left[\frac{(\omega_2-\omega_1)}{\pi}\right] d\omega' \qquad (14b)$$

Substitute Eq. (14b) into Eq. (13), change the limits of the integral accordingly, and simplify the integrand:

$$\bar{S}^2 = \frac{1}{(\omega_2-\omega_1)}\int_0^{\pi}[S(\omega(\omega'))]^2 \left[\frac{(\omega_2-\omega_1)}{\pi}\right] d\omega'$$

$$= \frac{1}{\pi}\int_0^{\pi}[S(\omega(\omega'))]^2 d\omega' \qquad (15)$$

Replace $S(\omega(\omega'))$ with its cosine expansion from Eq. (5) to obtain:

$$\bar{S}^2 = \frac{1}{\pi}\int_0^{\pi}[C_0' + \sum_{l=1}^{N} C_l' \cos(l\omega')]^2\, d\omega' \qquad (16)$$

Parseval's theorem applied to Eq. (16) yields the mean-square solution for BLCCs:

$$(\bar{S}^2)_N = (C_0')^2 + \frac{1}{2}\sum_{l=1}^{N}(C_l')^2 \qquad (17)$$

4.3.2. Error measure

As shown in Table 1, the BLCC series tends towards zero from MW up to the upper bound M selected for illustrative purposes. One way of quantifying this behaviour is to calculate the mean-square differences between the BLCC series extended as far as M and the same series truncated one cepstral coefficient at a time. These differences give an estimate of the truncation error E, and the formula adopted derives from Eq. (17) as follows:

$$E_l = (\bar{S}^2)_M - \sum_{l=0}^{M}(\bar{S}^2)_l \qquad (18)$$

4.3.3. Truncation error profiles for the two selected sub-bands

The BLCC analysis of the utterance "IOWA" was applied to all 97 frames of 25-msec duration (with 5-msec step size). The error measure E given in Eq. (18) was calculated at all frames and for the same sub-bands studied earlier.

Figures 4 and 5 display the (97-frame) means and standard deviations of E for the narrower and the wider sub-band, respectively. It is reassuring that the MW values based on all frames agree with those reported for the single frame considered in Section 4.1. More interestingly, there are two distinct regimes in the truncation error profiles. For both sub-bands, the largest error corresponding to the maximum peak of the mean curve is obtained by retaining only the zeroth-order BLCC. The means then become smaller and the spreads narrower as they turn into a flat regime of zero values. The turning point occurs near MW, beyond which the BLCC series may be assumed to contribute very little to the representation of the sub-band region.

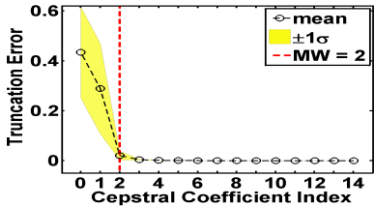


Figure 4: *Truncation error for the C_l': [0.1,0.9]-kHz extracted at all frames of "IOWA".*

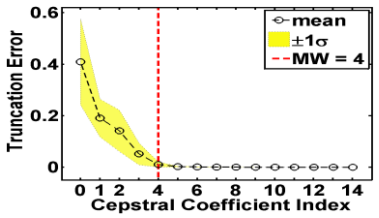


Figure 5: *Truncation error for the C_l': [2.5,4.0]-kHz extracted at all frames of "IOWA".*

5. Summary and application possibilities

We have uncovered a linear relationship, which permits direct transformation of the Fourier series of full-band cepstral coefficients into an analogous series of band-limited cepstral coefficients (BLCCs). These represent the spectral region of the full band delimited by the selected sub-band. The parametric expression of the linear transformation gives the flexibility of selecting the left and right limits for any sub-band within the frequency range of the full band.

We have also provided some empirical justification for truncating the BLCC series after MW without significant loss in spectral resolution. This is therefore the upper bound N proposed for practical use, although a value just larger than MW might conceivably be necessary for a more exact representation of the sub-band spectrum.

The flexibility and efficiency afforded by our band-limiting method suggest that the resulting BLCCs have the potential to throw new light on some challenging problems in speech science and technology. For instance, BLCCs could facilitate further systematic studies of the sub-band dependence of speaker [4,5,6] and accent [7] variability. They could also play a major part in the quest for robust classification performance. There is increasing evidence that the use of sub-bands is indeed beneficial for automatic speech [8] and speaker classification [9,10,11,12] and, more recently, for simulated [13,14] and real-world [15] forensic voice comparison.

Finally, it has not escaped our notice that the Euclidean distance between two sets of BLCCs is a corollary of the mean square formula in Eq. (17). For $N = MW$, we conjecture that the BLCC-based distance measure will approach the "exact" measure proposed in [16], which implicates the entire set of M full-band coefficients regardless of the sub-band's width. This conjecture points to the possibility of improving the efficiency and perhaps even the statistical stability of band-limited cepstral distances in speech and speaker classifiers.

6. Acknowledgements

I am grateful to Prof. Shunichi Ishihara for the uplifting discussions held locally in Canberra and remotely during his 2021-2022 sabbatical stay in Japan. I thank the anonymous reviewers for their suggestions. I extend sincere thanks to Prof. Phil Rose and Dr Yuko Kinoshita for their interest and queries.

7. References

[1] Deller Jr. J.R., Proakis, J.G. and Hansen, J.H.L., Discrete-Time Processing of Speech Signals, Macmillan, 1993.

[2] Picone, J.W., "Signal modelling techniques in speech recognition", *Proceedings of the IEEE*, 81(9): 1215-1247, 1993.

[3] Gradshteyn, I.S. and Ryshik, I.M., Tables of Integrals, Series and Products, Academic Press, 2007.

[4] Kitamura, T. and Akagi, M., "Speaker individualities in speech spectral envelopes", *Proc. 3rd International Conference on Spoken Language processing*, Yokohama, 1183-1186, 1994.

[5] Khodai-Joopari, M., Clermont, F. and Barlow, M., "Speaker variability on a continuum of spectral sub-bands from 297-speakers' non-contemporaneous cepstra of Japanese vowels", *Proc. 10th Australian International Conference on Speech Science and Technology*, Sydney, 504-509, 2004.

[6] Clermont, F., Kinoshita, Y. and Osanai, T., "Sub-band cepstral variability within and between speakers under microphone and mobile conditions: A preliminary investigation", *Proc. 16th Australasian International Conference on Speech Science and Technology*, Parametta, 317-320, 2016.

[7] Arslan, L.M. and Hansen, J.H.L, "A study of temporal features and frequency characteristics in American English foreign accent", *The Journal of the Acoustical Society of America,* 102(1): 28-40, 1997.

[8] Mokhtari, P. and Clermont, F., "Contributions of selected spectral regions to vowel classification accuracy", *Proc. 3rd International Conference on Spoken Language processing*, Yokohama, 1923-1926, 1994.

[9] Hayakawa, S. and Itakura, F., "Text-dependent speaker recognition using the information in the higher frequency band", *Proc. International Conference on Acoustics, Speech and Signal Processing,* Adelaide, 137-140, 1994.

[10] Finan, R.A, Damper, R.I. and Sapeluk, A.T., "Improved data modeling for text-dependent speaker recognition using sub-band processing", *International Journal of Speech Technology*, 4: 45-62, 2001.

[11] Sivakumaran, P., Ariyaeeinia, A.M. and Loomes, M.J., "Sub-band based text-dependent speaker verification", *Speech Communication*, 41: 485-509, 2003.

[12] Osanai, T., Kinoshita, Y. and Clermont, F., "Exploring sub-band cepstral distances for more robust speaker classification", *Proc. 17th Australasian International Conference on Speech Science and Technology*, Sydney, 41-44, 2018.

[13] Kinoshita, Y., Osanai, T. and Clermont, F., "Forensic voice comparison using sub-band cepstral distances as features: A first attempt with vowels from 306 Japanese speakers under channel mismatch conditions", *Proc. 17th Australasian International Conference on Speech Science and Technology*, Sydney, 45-48, 2018.

[14] Kinoshita, Y., Osanai, T., and Clermont, F., "Sub-band cepstral distance as an alternative to formants: Quantitative evidence from a forensic comparison experiment", 94, *Journal of Phonetics*, 2022.

[15] Rose, P., "Likelihood ratio-based forensic semi-automatic speaker identification with alveolar fricative spectra in a real-world case", *Proc. 18th Australasian International Conference on Speech Science and Technology*, Canberra, 2022.

[16] Clermont, F. and Mokhtari, P., "Frequency-band specification in cepstral distance computation", *Proc. 5th Australian International Conference on Speech Science and Technology,* Perth, I: 354-359, 1994.

Markedness in Kaytetye Reduplication: An Information-Theoretic Analysis

Forrest Panther[1], Mark Harvey[2]

[1]The New Zealand Institute of Language, Brain & Behaviour, University of Canterbury, New Zealand
[2]University of Newcastle, Australia
forrest.panther@canterbury.ac.nz; mark.harvey@newcastle.edu.au

Abstract

Kaytetye is an Arandic language spoken in Central Australia. Kaytetye has a partial reduplication pattern, in which the reduplicant is a disyllable. Surprisingly, the reduplicant in this form is required to be vowel-initial, and consonant-initial forms are ungrammatical. We show through quantitative analysis of the Kaytetye lexicon that the form of the reduplicant results from the distribution of information: vowel-initial forms are less informationally complex than consonant-initial forms, and an informational 'floor' effect constrains the minimal size. These results support an analysis of markedness that it results from information optimization within a language, not from universal patterns of markedness.

Index Terms: markedness, phonology, information, phonotactics, Australian languages, reduplication

1. Introduction

Markedness is defined as a factor that makes a construction or form more complex or rare, resulting in it being restricted in distribution, or even ungrammatical [1-4]. Hume [5] identifies two broad approaches to markedness in the literature:

- *Universal Markedness*: Markedness results from universal patterns of complexity and simplicity. Therefore, 'unmarked' and 'marked' patterns are identical for all languages. Markedness is a property of Universal Grammar, as framed in formal approaches to phonology [1, 3, 6].
- *Language-Specific Markedness*: Markedness results from communication optimization; in other words, it emerges from speaker-specific and language-specific phenomena [4, 7-11]. Markedness in phonology, in part, results from phonotactics and the distribution of information in that language, which affect these patterns [5, 12]

Under *Universal Markedness*, languages will conform to universal patterns of markedness, while under *Language-Specific Markedness*, markedness patterns emerge from the structuring of information in the specific language. This latter approach, of course, predicts recurring patterns across languages, as languages generally share many common properties, and the facts of human psychology and biology will result in similarities in markedness patterns across languages. However, it also predicts that there will be occasional patterns that are consistent with patterns in that language, but not with patterns common in language generally. Evidence of the existence (or non-existence) of these patterns allows these two analyses of markedness to be empirically testable.

2. Reduplication in Kaytetye

2.1. Markedness in Reduplication

Markedness is a factor in the formation of partial reduplicants [6]. For example, in Nookta (Wakashan, Pacific North-West), a reduplication pattern selects the initial consonant of a word and the following vowel: *čičims'iḥ* [13]. The reduplicant is the initial syllable of the base, apart from the coda of that syllable, i.e. the reduplicant is *či*, not *čim*. The analysis is that CV syllables are unmarked while CVC syllables are marked, as evidenced by syllable typology [10, 14].

Partial reduplicants also typically show a fixed size. In the Nookta reduplication pattern, the shape is CV; CVC is ungrammatical. These shapes typically correspond to "templates", such as a syllable, or a prosodic foot [15]. Earlier models of reduplication proposed that the size of the reduplicant is fixed and not reducible to any other factors [16-17]; some recent models have also maintained this general approach [18]. However, more recent approaches, such as Generalized Template Theory, view the size of the reduplicant as being constrained by markedness, just like other constraints on the content of the reduplicant [19-21].

2.2. Kaytetye Verbal Reduplication

Kaytetye verbal reduplication occurs in an Associated Path (henceforth AP) construction, a set of constructions that associate a path to a predicate [22-24]. Examples (1) & (2) illustrate a sentence that occurs with and without an Associated Path construction with the root /kwaṯə-/ 'drink'. When in an AP construction, as in (2), the verb /kwaṯə-/ receives a 'participial suffix' /j(ə)/, that indicates that the action (in this case drinking) occurs after the path. It is then followed by a 'path auxilliary' /alpə-/, which indicates a path back to a location, with the meaning of the overall construction meaning 'drink when getting back'.

1) /acə aɳtu kwaṯə-n̪ə/
1SG.ERG water drink-PST.PFV
"I drank water."

2) /acə aɳtu kwaṯə-j+**alpə**-n̪ə/
1SG.ERG water drink-after+**return**-PST.PFV
"I drank water when I got back."

Most Associated Path constructions have Path Auxiliaries that correspond to a lexical verb -- /alpə-/ means 'go back' (compare English 'have' as both a lexical verb and an auxiliary verb). However, the most frequent Associated Path construction has a reduplicant as the Path Auxiliary. In this construction, the verb

occurs with the participial suffix /-lp(ə)/, which means 'during'. The reduplicant indicates a path without any particular direction away or from a place, which is translated as '(on) the way', and consequently the meaning of the construction is 'do X on the way'.

The reduplicant in this Associated Path construction may take two forms. When the verb root in the construction is a consonant-initial monosyllable (henceforth CV), the reduplicant is a total copy of the verb root. A constructed example is shown in (3) involving the verb root /pu-/ 'cook'.

3) /ɻə a:rə pu-lpə+**pu**-n̪ə/
3SG.NOM kangaroo cook-during+**RED**-PST.PFV
"He cooked a kangaroo on the way."

When the verb root is larger than CV (i.e. at least disyllabic), the reduplicant has the shape of a vowel-initial disyllable (henceforth VCV). Constructed examples based on attested forms, are shown in (4)-(6).

4) /ɻə a:rə alarə-lp+**arə**-n̪ə/
3SG.NOM kangaroo kill-during+**RED**-PST.PFV
"He killed a kangaroo on the way."

5) /acə aŋʧu kwaṯə-lp+**aṯə**-nə/
1SG.ERG water drink-during+**RED**-PST.PFV
"He drank water on the way."

6) /ərmicinə akuwiɲcə-lp+**iɲcə**-n̪ə/
dust rise-during+**RED**-PST.PFV
"The dust rose."

In these constructions, the content of the reduplicant is copied from the right edge of the verb root base. Note, however, that the reduplicant does not copy the onset of the penultimate syllable, and to do so would produce an ungrammatical form. This requirement is typologically non-standard: if the shape of reduplicants is conditioned by markedness, then while they may show obligatory onsets, they will not typically show vowel-initial conditions. This is because syllables with onsets are analyzed as less marked than syllables without onsets [10, 14]. This then raises the question of whether there is a motivation for this pattern within Kaytetye itself, and by extension, the *Language-Specific Markedness* analysis, as opposed to the *Universal Markedness* analysis. The goal of this paper is to show that, in fact, this pattern is predicted under this *Language-Specific* approach.

2.3. Hypotheses

The *Language-Specific Markedness* analysis of Kaytetye reduplication hypothesizes that the forms are selected due to their lack of complexity. That is, CV and VCV forms are less complex than CVCV forms. We hypothesize that this follows from existing patterns in the Kaytetye lexicon.

The monosyllabic and disyllabic shapes are motivated by the minimal information content of short forms. However, we hypothesize that the predominance of disyllabic forms relates to an information 'floor' effect. While the shape of the reduplicant is informationally simple, it shows an informational minimality effect, in which any productive smaller form would be too simple to be informative as a reduplicative auxiliary verb.

The literature on gradient phonotactics agrees that speakers use the statistics of types in their mental lexicon to develop phonotactic intuitions [25-31]. Consequently, patterns relating to phonological complexity will be inferable from statistics in the Kaytetye lexicon. Based on the shape of the Kaytetye reduplicants, we have three hypotheses about the distribution of roots in the Kaytetye lexicon:

H1. There will be a positive correlation between: (i) the number of syllables; (ii) the initial phonotactics (i.e. whether the root is consonant or vowel-initial) of lexical items, and their overall complexity.
H2. The effect of initial phonotactics on the complexity of lexical items will be present independent of the number of segments overall in an item. In other words, the average complexity of the configurations in an item will still show the effect of initial phonotactics.
H3. There will be distributional evidence that monosyllabic roots are informationally deficient in a way disyllables are not.

3. Method

3.1. *kRoot*, a Database of Kaytetye Word Roots

In order to carry out an analysis of the Kaytetye lexicon, it is necessary to have a representative list of Kaytetye lexemes. Consequently, we developed a set of Kaytetye word roots from headwords in the *Kaytetye-to-English Dictionary* [32].

From the list of all the Kaytetye headwords, we removed any items that had their listed category as an affix or a clitic. We removed multi-word expressions, and forms with a hyphen. Headwords with no assigned part of speech were also removed. From the remaining set, we used replacement procedures using regular expressions to remove any transparent morphology, and the resulting set was hand corrected for any possible errors in the remaining data. From this remaining set, items that were judged to be word roots met one of two criteria: (i) there was no evidence of morphological complexity in the item; (ii) the item appeared complex, but its semantics were not transparently derived from these components. An example of the latter category is the word *akwerrepenhe* /akurəpən̪ə/, a word that appears to be transparently 'from the coolamon' (i.e. the word /akurə/ 'coolamon' in the sequential case), but means 'small baby'. Finally, semantically transparent roots in complex headwords that were not independent entries in the dictionary were added to the database as independent roots.

This process resulted in a set of 2,762 word roots belonging to seven parts of speech: noun, verb, pronoun, demonstrative, preverb, coverb, and adverb. The dataset was then transformed from Kaytetye orthography into an IPA representation using regular expressions. This IPA repesentation made use of the four-vowel phonemic analysis of [23].

3.2. Complexity Scores

The kRoot dataset was used to train a bigram model. For the modelling, root boundary characters (represented as '#') were added to indicate the beginning and the end of the root. The bigram in each root type was then retrieved. For example, for a root type /aləkə/, boundary characters were added: *#aləkə#*. This root then contributed six bigrams: *#a*, *al*, *lə*, *ək*, *kə*, *ə#*. This process was conducted for every root type in the database, and the full list of bigrams was retrieved.

The score calculated for each unique bigram is the positive log of its conditional probability, which is used in research to model the phonotactics of word forms [25, 28, 33-35]. The

conditional probability of the bigram was calculated as the frequency of the bigram in the set of roots in the database divided by the frequency of the initial unigram. For example, the conditional probability of *al* is the equivalent of the frequency of *al* divided by the frequency of *a*. When this conditional probability was retrieved, the negative base-2 log probability was derived to produce a "bits per phoneme" measure as an estimate for the complexity of the bigram. The formula used to calculate the score of the bigram *al* is summarized in (7).

7) $\text{score(al)} = \log_2(\frac{Count(al)}{Count(a)})$

We used this bigram model to calculate two measures of the complexity of word roots in the *kRoot* dataset, both of which were used in the data modelling. A *sum complexity score* is the sum of the complexity scores of the bigrams in a word root. The *average complexity score* is the mean complexity score of the bigrams in a word root.

3.3. Modelling Procedure

In order to determine the effect that root size and root-initial phonotactics have on the sum and mean complexity of Kaytetye word roots, we created two linear mixed effects regression models using the *lme4* package [36] in R [37]. **Model 1** uses the sum complexity of roots as the response variable, while **Model 2** uses the average complexity of roots as the response variable.

In developing these models, we considered two fixed factors: (i) number of syllables in the root (1 - 9 syllables); (ii) whether the root begins with a vowel (True/False). Random intercepts considered were: (i) the part of speech of the root; (ii) the quality of the final vowel (data exploration showed that this has a significant effect on the complexity of the root). Linear modelling utilizes the *kRoot* dataset, excluding 8 and 9 syllable roots due to the absence of consonant-initial forms at these sizes.

4. Results

Descriptive StatisticsTable 1 summarizes the distribution of the *kRoot* word roots by number of syllables and whether they are consonant-initial or vowel-initial. Roots of 8 and 9 syllables are excluded. Only 24 roots are monosyllables, which is less than 1% of the set of word roots. All but two of these are consonant-initial, and the two vowel-initial forms are homphones: /a:/, which can mean 'fight', or 'entrance'. The smallest size of root with a significant proportion of the lexicon is the disyllables, with 560 items, or 20.3% of the set of roots.

For every root size apart from monosyllables and disyllables, the number of vowel-initial forms is higher than the number of consonant-initial forms. Vowel-initial forms are quantitatively predominant overall: of the 2,762 roots in the *kRoot* dataset, 1,705 (61.7%) are vowel-initial.

Figure 1 provides the mean sum and average complexity scores for *kRoot* roots, categorized by their syllable count and root-initial phonotactics. The vowel-initial monosyllable category is excluded from this figure, which has only one root type (two roots with the form /a:/). It is, however, included in the modelling.

First, in support of Hypothesis H1, these results show that there is the expected positive correlation between the size of word roots and their sum complexity. Second, in support of

Syllable Count	Vowel-Initial	Consonant-Initial
1	2	22
2	217	343
3	756	458
4	507	182
5	162	47
6	47	2
7	11	3

Table 1: Distribution of Kaytetye Word Roots by Syllable Count and Root-Initial Phonotactics

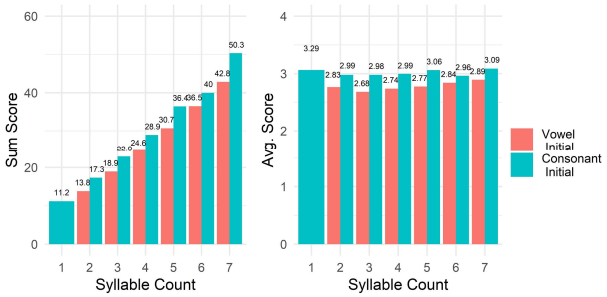


Figure 1: Mean Sum and Average Complexity Scores By Syllable Count and Root-Initial Phonotactics

Hypothesis H2, there is an effect of being vowel-initial on the average complexity in word roots. However, there is not an obvious effect of word root size in the average complexity score. The exception to this is in the monosyllable category, which shows the highest mean average complexity score of all the categories presented here. One possible interpretation of this is that it results from a combination of the small sample size of this category, and the low number of bigrams in comparison to other categories. In this instance, a comparatively small number of infrequent phonotactic configurations can inflate the score in this category, while in categories with more bigrams these configurations will be averaged out. This interpretation is unlikely given the fact that other categories with low numbers of bigrams (i.e. disyllabes) and categories with low counts (such as 7-syllables) show numbers consistent with the rest of the average complexity counts. Instead, it is likely these numbers reflect more on-average complex configurations in monosyllables. This possibility is supported by other aspects of the distribution of monosyllabic roots, that point to a preference for contexts with more information. That is, a vast majority of monosyllabic forms either occur in contexts where there is morphology to provide context for their interpretation, or they are function words that are frequent in discourse. Of the 24 monosyllables, 12 (50%) are verbs, which in Kaytetye always occur with a verbal suffix. Only four (20%) are nouns, and two of these nouns are homophones that consist of the rare long vowel /a:/. Four CV roots are pronouns, and two are demonstratives – both are closed classes. The other two are coverbs, which always occur with a following verb. In other words, monosyllables are over-represented in classes where there is context for their interpretation. Compare this distribution with disyllables, in which 335 (59.8%) are nouns and only 112 (20%) are verb roots.

All these observations are consistent with the notion of a type of complexity 'floor', in which forms below this floor show patterns that increase their salience. These observations support Hypothesis H3.

Model 1				
Fixed Effect	Estimate	*t* value	*p* value	Sig.
Intercept	2.5	1.2	0.42	
Initial Segment – Cons.	3.5	7.9	<0.001	***
Syllables	5.9	70.4	<0.001	***
Init. Segment – Cons. : Syl.	0.3	1.9	0.054	
Model 2				
Fixed Effect	Estimate	*t* value	*p* value	Sig.
Intercept	2.9	8.5	0.072	
Initial Segment – Cons.	0.4	8.9	<0.001	***
Syllables	0.05	5.4	<0.001	***
Init. Segment – Cons. : Syl.	-0.03	-2.2	<0.05	*

Table 2: Summary of Fixed Effects for Model 1 (Sum Complexity) & Model 2 (Average Complexity)

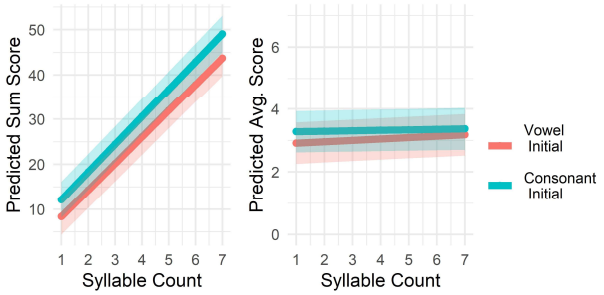


Figure 2: Predicted Sum and Average Complexity Score in Model 1 (left) & Model 2 (right) by Syllable Count and Initial Phonotactics.

4.1. Model Results

In order to confirm the findings relating to Hypothesis H1 and Hypothesis H2 in the distribution of complexity scores, we conducted data modelling on the sum (Model 1) and mean (Model 2) complexity scores of word roots. The resulting models included syllable count and whether the root is vowel-initial as fixed effects, including an interaction term between these two factors.

A summary of the fixed effects of Model 1 & Model 2 are shown in Table 2. In all, Model 1 shows the expected positive effect of vowel-initial forms and syllable count on the sum complexity score, meeting the predictions of Hypothesis H1. Model 2 shows the effect of vowel-initial roots on average complexity scores, confirming that the predictions of Hypothesis H2 are met. Surprisingly, Model 2 also shows an effect of syllable count, as well as an interaction between vowel-initial roots and syllable count. This effect is observable in Figure 2: vowel-initial roots show a positive correlation between average complexity score and syllable count, while this trend is almost non-existent for consonant-initial roots. We propose that this effect relates to the comparatively high complexity of consonant-initial monosyllables, that was described above. In other words, the comparatively high complexity of monosyllables largely negates the otherwise apparent trend of increasing average complexity of roots by their syllable count. These statistical results confirm the predictions of Hypothesis H1 & H2, and are consistent with the prediction of Hypothesis H3.

5. Discussion

The results presented in this paper show that the form of the Kaytetye verbal reduplicant is predictable based on the information content of the Kaytetye lexicon. The vowel-initial condition is based on the fact that consonant-initial forms are more informationally complex than vowel-initial forms. The disyllabic default of the reduplicant is based on the low information content of monosyllables, showing an apparent 'floor' effect. This is consistent with a *Language-Specific Markedness* analysis of markedness.

The focus of this paper has been presenting evidence in favour of a *Language-Specific Markedness* analysis to markedness in Kaytetye reduplication, rather than presenting evidence against the *Universal Markedness* analysis. However, we noted briefly in §2.2 that the vowel-initial condition is inconsistent with the evidence of markedness and syllable typology, which favours consonant-initial forms rather than vowel-initial forms [10, 14]. VC reduplicants appear in other languages. For example, Lushootseed exhibits a reduplication pattern with a VC shape: ʔəxid → ʔəx-**ɪx**-əd [20]. However, the reduplicant attaches after C2 in the base, and this C2 provides an onset for the reduplicant when the overall Base + Reduplicant word is syllabified. Consequently, the universal preference for onsets is satisfied.

This might also appear a plausible analysis of the Kaytetye reduplication. For example, in the construction in (4) alarə-lp+**arə-ɲə** 'kill-during+**RED**-PST.PFV', the participial suffix could appear to be underlyingly consonant-final, i.e /-lp/. If it was underlyingly consonant-final, then this could license a following V-initial reduplicant because the participial suffix would be available to supply an onset. However, all morphemes in Kaytetye are V-final, and the final vowel of the participial suffix surfaces when there is a following C-initial morpheme, as in (8).

8) /acɳu-lpə=lk+acɳu-jə ɻə/
descend-during=then+RED-FUT 3SG.NOM
'Then he will go down.' (Example from [32])

The appearance of the consonant-final form of the participial suffix is due to vowel hiatus resolution, which deletes the first of two adjacent vowels, and occurs in other aspects of Kaytetye phonology [24]. Thus, (4) is underlyingly /alarə-lpə+arə-ɲə/, but the hiatus sequence /ə+a/ between the participial suffix and the reduplicative auxiliary root is not permitted. Hiatus reduction is a post-lexical, rather than a lexical process, and so will not satisfy onset maximization at a lexical level.

In this paper we showed that the facts of Kaytetye verbal reduplication are explicable under a *Language Specific Markedness* analysis, while there are significant problems with a *Universal Markedness* analysis. These results necessarily raise questions for the motivations for reduplication patterns generally, and whether alternative approaches are necessary.

6. References

[1] Rice, K. (2007). Markedness in phonology. In P. De Lacy (Ed.). Cambridge University Press.

[2] Hume, E. (2011). Markedness. In M. van Oostendorp, C. J. Ewen, E. Hume, & K. Rice (Eds.), The blackwell companion to phonology. Wiley-Blackwell. https://doi.org/10.1002/9781444335262

[3] Kager, R. (1999). Optimality theory. Cambridge university press.

[4] Hume, E. (2016). Phonological markedness and its relation to the uncertainty of words. Phonological Studies. Journal of the Phonological Society of Japan, 19, 107–116.

[5] Hume, E. (2006). Language specific and universal markedness: An information-theoretic approach. Linguistic Society of America Annual Meeting: Colloquium on Information Theory and Phonology.
[6] McCarthy, J., & Prince, A. S. (1994). The emergence of the unmarked: Optimality in prosodic morphology. In M. Gonzàlez (Ed.), Proceedings of the north east linguistic society 24.
[7] Hume, E. (2003). Language specific markedness: The case of place of articulation. Studies in Phonetics, Phonology & Morphology.
[8] Hume, E. (2008). Markedness and the language user. Phonological studies, 11.
[9] Comrie, B. (1986). Markedness, grammar, people, and the world. Markedness (pp. 85–106). Springer.
[10] Blevins, J. (2006b). Syllable: Typology. In K. Brown (Ed.), Encyclopedia of language and linguistics (second). Elsevier.
[11] Bybee, J. (2010). Markedness: Iconicity, economy, and frequency. In J. J. Song (Ed.), The oxford handbook of linguistic typology. Oxford University Press.
[12] Hume, E. (2004). Markedness: A predictability-based approach. Annual Meeting of the Berkeley Linguistics Society 30: General Session and Parasession on Conceptual Structure and Cognition in Grammatical Theory, 182–198. https://doi.org/10.3765/bls.v30i1.948
[13] Stonham, J. T. (1990). Current issues in morphological theory (Doctoral dissertation). Stanford University.
[14] Zec, D. (2007). The syllable. In P. De Lacy (Ed.), The cambridge handbook of phonology (pp. 161–194). Cambridge University Press.
[15] McCarthy, J., & Prince, A. S. (1995). Faithfulness and reduplicative identity. Papers in Optimality Theory, 10. https://doi.org/10.7282/T31R6NJ
[16] McCarthy, J. (1979). Formal problems in semitic phonology and morphology (Doctoral dissertation). MIT. Cambridge, MA.
[17] Marantz, A. (1982). Re reduplication. Linguistic inquiry, 13(3), 435–482.
[18] Saba Kirchner, J. (2010). Minimal reduplication (Doctoral dissertation). University of California, Santa Cruz.
[19] Urbanczyk, S. (1996). Morphological templates in reduplication. North East Linguistics Society, 26(1).
[20] Urbanczyk, S. (2006). Reduplicative form and the root-affix asymmetry. Natural Language & Linguistic Theory, 24(1), 179–240.
[21] McCarthy, J., & Prince, A. S. (1986). Prosodic morphology [Unpublished Manuscript, University of Massachusetts, Amherst & Brandeis University].
[22] Koch, H. (1984). The category of 'associated motion' in kaytej. Language in central Australia, 1(1), 23–34.
[23] Panther, F., & Harvey, M. (2020). Associated path in Kaytetye. Australian Journal of Linguistics, 40(1),74–105. https://doi.org/10.1080/07268602.2019.1703644
[24] Panther, F. (2021). Topics in Kaytetye phonology and morpho-syntax (Doctoral dissertation). University of Newcastle
[25] Coleman, J., & Pierrehumbert, J. (1997). Stochastic phonological grammars and acceptability. In J. Coleman (Ed.), Computational phonology. third meeting of the ACL special interest group in computational phonology (pp. 49–56).
[26] Vitevitch, M. S., & Luce, P. A. (1999). Probabilistic phonotactics and neighborhood activation in spoken word recognition. Journal of memory and language, 40(3), 374–408.
[27] Hay, J., Pierrehumbert, J., & Beckman, M. (2004). Speech perception, well-formedness and the statistics of the lexicon. Phonetic interpretation: Papers in laboratory phonology vi (pp. 58–74). Cambridge University Press.
[28] Vitevitch, M. S., & Luce, P. A. (2004). A web-based interface to calculate phonotactic probability for words and nonwords in english. Behavior Research Methods, Instruments, & Computers, 36(3), 481–487.
[29] Frisch, S. A., Large, N. R., Zawaydeh, B., & Pisoni, D. B. (2001). Emergent phonotactic generalizations in english and arabic. In J. Bybee & P. Hopper (Eds.), Frequency and the emergence of linguistic structure. John Benjamins.
[30] Richtsmeier, P. T. (2011). Word-types not word-tokens, facilitate extraction of phonotactic sequences by adults. Laboratory Phonology, 2, 157–183.
[31] Oh, Y., Todd, S., Beckner, C., Hay, J., King, J., & Needle, J. (2020). Non-Māori-speaking new Zealanders have a Māori proto-lexicon. Scientific reports, 10(1), 1–9.
[32] Turpin, M., & Ross, A. (2012). Kaytetye to English dictionary. IAD Press.
[33] Goldsmith, J., & Riggle, J. (2012). Information theoretic approaches to phonological structure: The case of Finnish vowel harmony. Natural Language & Linguistic Theory, 30(3), 859–896.
[34] Pimentel, T., Roark, B., & Cotterell, R. (2020). Phonotactic complexity and its trade-offs. Transactions of the Association for Computational Linguistics, 8, 1–18.
[35] Brown, P. F., Della Pietra, S. A., Della Pietra, V. J., Lai, J. C., & Mercer, R. L. (1992). An estimate of an upper bound for the entropy of English. Computational Linguistics, 18(1), 31–40.
[36] Bates, D., Maechler, M., Bolker, B., & Walker, S. (2021). Package 'lme4'. R package version.
[37] R Core Team. (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing. Vienna, Austria.

Warlpiri IDS: Expanding the path to communicative success

Rikke Bundgaard-Nielsen[1,2], Carmel O'Shannessy[2], Alice Nelson[3], Jessie Bartlett[3], Vanessa Davis[4]

[1]MARCS Institute for Brain, Behaviour and Development; Western Sydney University; [2]School of Literature, Languages and Linguistics, Australian National University; [3]Red Dust Role Models;

[4]Tangentyere Council Research Hub

rikkelou@gmail.com; carmel.oshannessy@anu.edu.au; alice@reddust.org.au; jessie@reddust.org.au; vanessa.davis@tangentyere.org.au

Abstract

The present study examines the spectral and temporal characteristics of Infant Directed Speech (IDS) vowels spoken to young children in the Indigenous Australian language Warlpiri. The results show that vowel hyperarticulation, temporal expansion and pitch raising are characteristics of Warlpiri IDS to children in the third year of life. The results also suggest that vowel space fronting may be a common strategy, and that vowel space and durational expansion may play a didactic role in IDS young children at an age characterised by rapid vocabulary expansion and increased multiword utterances
Index Terms: IDS, ADS, Warlpiri, Vowels, Nouns.

1. Special Speech Registers for Children

The use and characteristics of special Infant and Child Directed Speech registers (IDS; CDS; or in older literature Baby Talk; Motherese) has been argued to be a supportive strategy adopted by carers (mothers, fathers, other carers, and older children) with the aim to scaffold language acquisition. Questions of the universality of IDS have received significant attention in the literature, and cross-linguistic research has demonstrated differences across languages and cultures, and on changes in the characteristics of IDS across development. The present study examines the spectral and temporal characteristics of IDS vowels spoken to young children in the Indigenous Australian language Warlpiri. The results show that vowel hyperarticulation, temporal expansion and pitch raising are characteristics of Warlpiri IDS to children in the early multiword stage of language acquisition and beyond. The results also suggest that vowel space fronting may be a common strategy observed in Warlpiri IDS, and that vowel space and durational expansion may play a particular didactic role in the speech to young children who are at an age where rapid vocabulary expansion is typical, as nouns appear to be the locus of the greatest IDS modifications.

IDS is characterised by a slower speech rate, higher fundamental frequency, and greater pitch variations [1], longer pauses, often repetitive intonational structures [2], and shorter sentences with a more limited lexicon than Adult Directed Speech (ADS) [3]. IDS is also characterised by more hyperarticulation: vowels (and consonants) are often given articulatory/acoustically more extreme realisations, resulting in an expanded articulatory/acoustic vowel space, and increased acoustic/articulatory differentiation [4]. Such vowel hyperarticulation is also a feature of Foreigner Directed Speech (FDS: [5], but not Pet Directed Speech (PDS: [6] unless the pet is a parrot [7], suggesting that slow and clear speech is used as a didactic strategy in communication with individuals/entities who are perceived to be language learners (or at least capable of some language learning) but who are still learning.

IDS has been hypothesised to serve a number of different functions, most likely simultaneously and to varying degrees at different developmental stages, including **(1)** regulating infant attention [8], [9], [10]; **(2)** communicating affect and supporting social interaction [11], [12], [13], and; **(3)** supporting language acquisition [14], [15], [16], [17], the latter of which is the focus here. Vowel hyperarticulation in IDS has been particularly highlighted as potentially facilitating certain aspects of language acquisition: in particular, *segmental acquisition* (the learning of vowels and consonants), and in word-learning. In particular for the first year of life, vowel hyperarticulation has been argued to enhance *segmental learning*, as argued by [18] by providing infants with input containing high-quality and maximally differentiated vowel tokens that infants might attend to preferentially, perhaps due to its prosodic characteristics, as discussed above. Not all studies have shown that all vowels are uniformly *hyper*articulated: In at least one study [19], mothers were found to *hypo*articulate back vowels. This was interpreted to make the articulation more visibly accessible to infants than they are in ADS and thus rather than representing a goal of target undershoot, it indicated *enhancement* in the visual domain (as opposed to the acoustic).

In the second year of life, vowel hyperarticulation has been argued to be helpful in terms of *word-learning*, with research showing that the degree of vowel hyperarticulation in the maternal input at 18 months of age is correlated with the size of the receptive and productive language at 24 months of age [20]. This suggests that extra clarity in the phonemic specifications of words in IDS supports the acquisition of new vocabulary items. Other research has shown that IDS increases neural activity in 6- and 13-month-old infants compared to ADS, which again is argued to assist with the 'word spotting' of young word learners [21]. It is possible that increased neural activity also reflects recognition of being a potential addressee of IDS utterances [22], particularly in older infants and children. The findings in [20] are, however, also consistent with results showing that IDS vowel hyperarticulation (but not an enhanced pitch range) plays an important role in word recognition and word learning. And in one word recognition study, vowel hyperarticulation improved 19-month-olds' performance in word recognition tasks [23]. In another study, using a word learning paradigm, [24] showed that 21-month-olds learned new words only in IDS unless they already had large vocabularies, while 27-month-old toddlers learned new words in both ADS and IDS.

The acoustic characteristics of IDS, however, are not stable across early development [25], [26], [27]. This is generally taken to indicate that carers finetune their IDS to the

developmental stage of the child, including its linguistic development. In a cross-sectional design, [28] showed that the use of vowel space expansion and vowel target precision in IDS changes from infancy into toddlerhood, such that vowels are more hyperarticulated to early word-learners than prelingual infants, and young children who are combining words receive input with particularly hyperarticulated vowels—in both content words and function words. Longitudinal studies have also demonstrated that IDS segmental modifications (hyperarticulation and avoidance of segmental reduction patterns found in adult speech) changes over time. In another study, mothers were found to hyperarticulate vowels to pre-linguistic (<12 months of age) much more than to 5-year-old children and to adults [26], and [25] show that mothers' use of segmental deletion increases between 1;6 and 2;0 years of age only to decrease again between 2;0- 2;6 years of age, suggesting deliberate avoidance of deletion in speech to children during segmental acquisition and again in the early multiword stage of language acquisition. It is plausible that these changes in input characteristics reflect a 'first wave' of hyperarticulation to support vowel learning, and a 'second wave' to support rapid word-learning in the second year of life.

The research findings reported above, however, are predominantly based on studies undertaken in just a few of the worlds' languages [29], leaving large gaps in our knowledge of cross-linguistic IDS patterns. And while most studies indicate that carers do use special registers with children, not all studies have found that carers hyperarticulate vowels at all [30], [31]. Likewise, crosslinguistic comparisons have demonstrated differences in the implementation or degree of hyperarticulation and pitch raising across different languages, when hyperarticulation is present (e.g., [32]). The prosodic system of the target adult language has also been found to influence the particular prosodic modifications made in IDS ([33], as do sociolinguistic considerations ([34], [35]). Taken together, this suggests that the 'shape' of IDS in each language is subject to significant variability and reflects linguistic and sociolinguistic demands outside of the carer-child interaction and the developmental characteristics of the child.

Warlpiri is a Pama-Nyungan language spoken by approximately 2500-3000 people, mostly in the remote communities of Yuendumu, Lajamanu, Nyirrpi, and Willowra, in the Northern Territory (NT), and in regional towns and cities, including Alice Springs and Darwin, NT. Although endangered, Warlpiri continues to be learned by children as their primary language in three of the communities (Yuendumu, Nyirrpi, and Willowra), and as one of their first languages in Lajamanu. Families may travel between the regional towns and cities and remote communities and spend varying amounts of time in one location. The families in this study were recorded while in Alice Springs, where they were staying temporarily.
Phonologically, Warlpiri is characterized by a single series of stops /p t ʈ c k/ with five main places of articulation. This is repeated in the nasal series /m n ɳ ɲ ŋ /, and Warlpiri further has three laterals /l ɭ ʎ/, two approximants /w j/, and two or three rhotic phonemes; trill /r/ approximant /ɻ/, and a retroflex flap /ɽ/, though the phoneme status of the latter has been recently questioned [36]. In terms of vowels, Warlpiri sports only three: /i/, /a/, and /u/, with a phonemic length contrast. The Warlpiri formant space is compact, with a relatively compressed F1 range [37]; who reports the F1/F2 values of a single female Warlpiri speaker, indicating an F1 range of appx. 470-600 Hz (/i/ and /u/ vs /a/, and an F2 range of 1200-2400 Hz (/u/ vs /i/). The three Warlpiri vowels are distributed unevenly in the lexicon according to a vowel count of all entries in PanLex (https://vocab.panlex.org/wbp-000?page=0), with /a/ contributing 45% of vowels, /i/ vowels contributing 33%, and /u/ only 22%, making /u/ only half as frequent as /a/. This is typical for many Australian Indigenous languages. The existing literature on the characteristics of 'Baby Talk' in Warlpiri (and the limited literature on IDS in other Australian languages) has attended particularly to the consonants and to changes to adult wordforms. [38] identifies patterns of substitution of coronal consonants (stops /t ʈ/, nasals /n ɳ/, and laminals /l ɭ/) with the corresponding lamino-palatal consonants /c ɲ ʎ), and the use of special IDS wordforms, for instance, 'apa' for 'ngapa' (water). [39] report similar segmental changes in IDS in the neighbouring language Gurindji Kriol. In addition to coronal place neutralisation through palatalisation, like in Warlpiri, [39] list patterns of rhotic replacement, cluster simplification, deletion, consonant harmony, and replacement of apical postalveolars with apical alveolars as characteristic of Gurindji Kriol IDS. The authors further highlight that these segmental modifications are very common cross-linguistically in IDS and in children's speech development. The present study does not examine consonant modification. Consonant substitution and cluster simplification in Warlpiri and Gurindji Kriol IDS potentially arise from a desire to create wordforms that children can more easily copy in their own production, or that reflect typical child-language forms ⌈38], ⌈39]. IDS phoneme substitution and coronal neutralisation through palatalisation may also be intended to aid speech perception and word recognition, though this of course may in some sense be a double-edged sword. The observed pattern of coronal neutralisation, for instance, effectively eliminates a series of consonant contrasts that can be difficult to discriminate even for adult speakers [40]. Substitutions may also increase the risk of phonological overlap in lexical items used by children. Unfortunately, there is currently no systematic information about the acquisition ages—in terms of speech production as well as speech perception—of the typologically unusual inventories of many Indigenous Australian languages, though cross-linguistic comparisons [41] and informal observations [38], [39] may give some indication of what we might expect for both production and perception. The existing literature on IDS in Warlpiri and other Australian languages also does not offer information about the potential systematic modification of IDS in terms of prosody, vowel quality and vowel quantity, the foci of the present study.

In the following, we report on the acoustic characteristics of vowels in IDS to young Warlpiri-acquiring children by comparing ADS and IDS vowel quality, quantity, and pitch within a group of three adult speakers. The study provides a first examination of IDS vowels in a three-vowel system, where pressures for contrast enhancement might differ from those in languages with more crowded vowel inventories, such as English; or vowel systems (such as Japanese) where vowel duration is phonemic (vowel length is contrastive in only in a very restricted set of words in Warlpiri: [42]). We also examine the characteristics of vowels in IDS nouns relative to 'general' IDS, testing the possibility that IDS at that stage is (at least in part) a didactic strategy for scaffolding vocabulary development in young word-learners.

2. Method

2.1. Participants

We report on data from two Warlpiri-speaking women (A and M) and one Warlpiri-speaking man (G), who were recorded

(video, audio) at their homes in Alice Springs, NT. The women were recorded interacting with each other, a third woman, and three young children, while taking part in play, storytelling and discussion activities, centered on the day-to-day child-rearing activities and interactions with the children. At the time of the recordings, A was in her 50s, and M was appx. 30 y.o. Two of the three children were 28 and 30 months at the time of the recording; the date of birth of the third is not known, but the child was approximately 24 months. The women and children are all close family. The male participant (G; 50+ y.o.) was recorded while taking part in a play and conversation activity with a young child (40 months) and the child's mother, all related.

2.2. Materials

The video/audio recordings of the three adult participants (A, M, G) were transcribed, glossed, and translated into English, and coded for the intended addressee (IDS or ADS on the basis of close viewing of the video recordings) in *ELAN 6.3*. Target vowels were then hand-segmented and labelled in *Praat 6.2.12*., and vowel duration, F0, F1, and F2 extracted using an automatic script. We extracted target measurements from as many IDS vowels as possible from each of the female participants, as well as a roughly matching number of ADS vowels. The male participant provided ~240 IDS vowels but had only one brief ADS interaction during the recording session, providing just nine /i/, /a/, and /u/ tokens. We further coded all IDS vowels from concrete nouns in the IDS of the three speakers, creating two subsets of IDS vowel data: Vowels from (concrete) nouns (IDS Nouns) and vowels from everywhere else (IDS Non-Nouns). Vowels degraded by environmental noise, overlapping talkers, etc., were excluded from the dataset. In all individual datasets, /a/ was vastly overrepresented relative to /i/ and /u/, consistent with the general distribution in Warlpiri. The distribution of vowels by speaker and vowel quality is presented in *Table 1*.

Table 1. *Summary of the number of vowels extracted by speaker (A, M, G) and condition (ADS, IDS; IDS Nouns, and IDS Non-Nouns).*

ID	Style	Total	/a/	/i/	/u/
A	ADS	320	196	71	53
	IDS	382	206	61	109
	IDS Nouns	120	45	30	45
	IDS Non-Nouns	262	164	34	64
M	ADS	189	115	50	24
	IDS	149	93	31	25
	IDS Nouns	57	31	14	12
	IDS Non-Nouns	92	62	17	13
G	ADS	9	5	2	2
	IDS	242	144	30	68
	IDS Nouns	84	45	13	26
	IDS Non-Nouns	158	99	17	42

Data extracted in a naturalistic setting, such as the data reported on here, is often less controlled than data collected under laboratory conditions, especially in terms of the vocabulary used, and the number of instances of the target vowels that can be extracted. This is unfortunate, but it is likely that such datasets have other advantages in terms of ecological validity. We also wish to highlight another difference between lab-based and 'naturalistic' data collection such as this. In many studies of vowel hyperarticulation in IDS, parents/caregivers are provided with special toys to induce the use of corner vowels /i/, /a/, and /u/ in the play interactions with the children. Parents are often also asked to discuss these objects with researchers in a separate (consecutive) session, to match the IDS dataset. In languages with large vowel systems such as English and Danish, it may make good practical sense to ensure sufficient target vowels are produced in each data collection session through this type of experimental manipulation, but the data collected in such a context may consequently reflect additional linguistic phenomena to those intended. Task-induced use of contrastive, focus and question intonation, for instance, may induce changes in pitch, vowel quantity, and vowel quality, without these phenomena reflecting aspects particular to IDS (or which may be overrepresented in the data).

3. Results

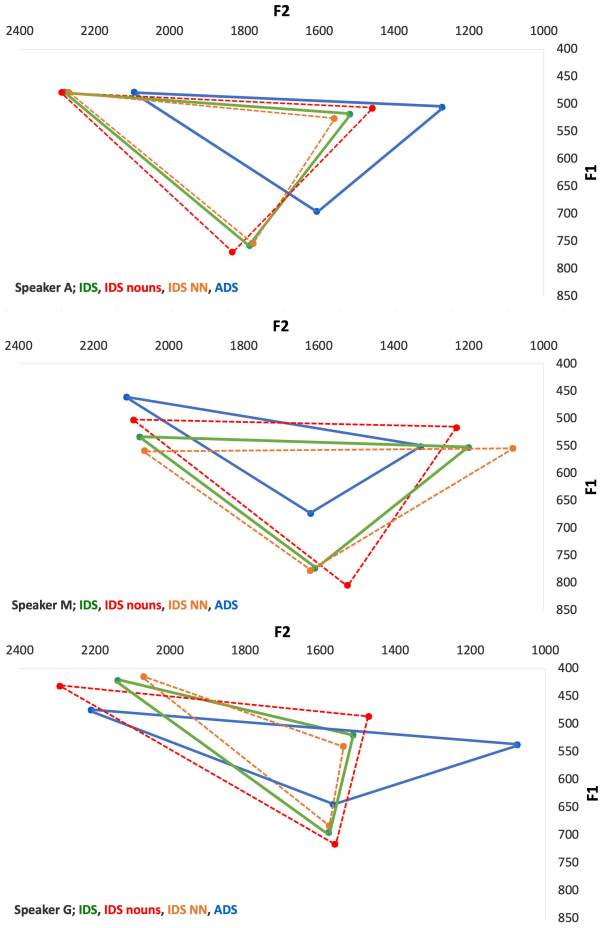
To assess whether vowel hyperarticulation, vowel expansion and pitch raising are characteristic of Warlpiri IDS, we compared vowel quality (F1, F2), vowel duration, and F0, in IDS and ADS for each of the three speakers separately. We present vowel duration and F0 in *Table 2* and F1/F2 plots for each speaker separately in *Figures 1-3*. Multivariate ANOVAs applied to *Speaker A* and *M's* vowel data revealed significant differences between their IDS and their ADS. *Speaker A's* IDS vowels were all longer and had a higher F2 than her ADS vowels (fronting), and in the case of /a/, also a higher F1 (lowering). *Speaker A*'s IDS /a/ also had a higher F0 than the ADS counterparts. *Speaker M*'s IDS /i/ and /a/ vowels were produced with a higher F1 (fronting) and a higher F0 than their ADS counterparts; but there were no significant differences in /u/, and no durational differences for any vowel. We did not conduct a statistical analysis of *Speaker G*'s data due to the low number of ADS observations, but we note a pattern of F2 and F1 raising (vowel fronting and lowering, in particular of /a/) in his dataset largely consistent with the two female speakers, and *Speaker A* especially. We also observe IDS vowel durations of almost double those in ADS, and likewise higher F0 (/a/, /u/).

Table 2. *Vowel duration for /a/, /i/, and /u/ (ms) and F0 (Hz) in ADS and IDS across speakers (A, M G).*

ID	Style	/a/ ADS/IDS	/i/ ADS/IDS	/u/ ADS/IDS
A	Dur.	72/94 ms	77/147 ms	60/100ms
	F0	203/214 Hz	206/218 Hz	207/206 Hz
M	Dur.	85/91 ms	89/105 ms	75/95 ms
	F0	197/230 Hz	200/220 Hz	203/226 Hz
G	Dur.	53/87 ms	51/84 ms	50/91 ms
	F0	195/216 Hz	218/209 Hz	157/216 Hz

We conducted a further analysis to investigate the hypothesis that IDS is a dynamic didactic modification, suited to the developmental needs of an infant/child. In the present study, the children were all in the early multiword stage of language acquisition, and we hypothesise that carer modifications may (among other things) be particularly targeted to word learning, and perhaps in particular to the teaching of 'names for things', consistent with the results of [20]-[24]. If that is the case, we would expect vowel hyperarticulation and temporal expansion to be driven by the characteristics of nouns rather than reflect IDS in general. We test this hypothesis by comparing vowel quality, duration and pitch in vowels extracted from IDS (concrete) nouns and vowels extracted from all other IDS words for each of the three speakers, again using

a series of Multivariate ANOVAs (See *Figure 1-3* for F1/F2 values, and *Table 3* for Duration and F0 values). In this analysis, we include Speaker G. The results indicate that /a/ and /i/ are longer in IDS nouns than in general IDS for *Speaker A*; /a/ is produced with a higher F2 and longer duration in IDS nouns produced by *Speaker M*, whose IDS Noun /u/s are also longer. *Speaker G* produces longer IDS Noun /a/s than in general IDS, and /i/s in Nouns that have a higher F1 than in IDS in general.



Figures 1-3: *F1-F2 vowel plots for Speakers A, M, G; IDS in green; ADS in blue lines. IDS Nouns and Non-Nouns (NN) in red and orange dotted lines.*

Finally, we assessed the degree of vowel space expansion in IDS versus ADS, and in IDS Nouns and IDS Non-Noun materials by calculating the Euclidian space enclosed by the F1/F2 values of each of the three vowels /i/, /a/, and u/. The ranking in terms of the vowel triangle of each vowel dataset is presented in *Table 4*. Minor idiosyncrasies aside, the data from the three speakers shows that IDS nouns are characterized by a much larger vowel space than IDS in general and IDS non-noun material, and that ADS tends to be characterized by the smallest vowel space of the four datasets. The exception—the relatively large vowel space characterising the ADS of *Speaker G*—must be taken with caution due to a very small dataset (See *Table 1*). The question of vowel space expansion is particularly interesting in the present dataset where two of the three speakers demonstrate substantial vowel space fronting, perhaps as argued by [19] to enhance the visual speech information, but this fronting does not result in vowel space reduction. It is also possible that the modifications observed bring the adult F1/F2 values closer to those observed in child speech and which children may prefer [43].

Table 3. *Vowel duration (ms) and F0 (Hz) for /a/, /i/, and /u/ in IDS Nouns (Ns) and IDS Non-Nouns (NNs) across speakers (A, M, G).*

ID	Style	/a/ Ns/NNs	/i/ Ns/NNs	/u/ Ns/NNs
A	Dur.	137/82 ms	117/172 ms	130/79 ms
	F0	221/212 Hz	215/221 Hz	206/206 Hz
M	Dur.	109/81ms	132/84 ms	135/59 ms
	F0	236/232 Hz	206/232 Hz	240/211 Hz
G	Dur.	106/79 ms	100/74 ms	109/79 ms
	F0	216/220 Hz	209/201 Hz	216/215 Hz

Table 4. *IDS, IDS Nouns (Ns), IDS Non-Nouns (NNs) and ADS ranked in terms of the Euclidian Space denoted by the F1/F2 values of vowels /i/, /a/ and /u/.*

ID	1st	2nd	3rd	4th
A	IDS Ns	IDS	IDS NNs	ADS
	△114022	△96300	△85788	△82599
M	IDS NNs	IDS Ns	IDS	ADS
	△108411	△101550	△100804	△79532
G	IDS Ns	ADS	IDS	IDS NNs
	△96851	△76211	△58361	△40413

4. Discussion

The study reported here is the first investigation of the acoustic characteristics of vowels in IDS in an Australian Indigenous language. The results show that adult speakers of Warlpiri raise the pitch, increase vowel duration, and produce more extreme F1/F2 values in their speech to young children compared to speech to (well-known) adults. This is largely consistent with what has been reported for IDS in languages from other parts of the world, though the strong indication of vowel space fronting in the data of two of the three speakers in the present study is not broadly attested, if arguably helpful in terms of improving visual speech information [19] or attracting the attention of children [43]. The study also indicates that hyperarticulation in Warlpiri IDS may serve a didactic purpose: The IDS analysed here was directed to young children who were rapidly acquiring new vocabulary and beginning to use multiword phrases, and the results from the present study clearly show that IDS Nouns stand out in terms of their acoustic characteristics: they have longer and more extreme vowels than other words in IDS. We suggest that this is consistent with patterns of avoidance of lenition processes [25] in speech to children of a similar age: reducing speech clarity is simply not helpful at an age where clarity of speech may assist a child to correctly learn and recognise new words. Interestingly, vowels from IDS nouns are not characterised by a higher F0 than Non-Noun IDS vowels, suggesting that F0 is not used for the same didactic purposes as vowel hyperarticulation and expansion. This appears consistent with reports that IDS pitch modulations are used to convey affect to young infants, and perhaps convey a range of emotions, or what we might call 'caregiver stance', to slightly older children, including indications of what behaviours are desirable and which are not, rather than assisting with word-learning. Finally, the results reported here suggest that studies of the acoustic characteristics that focus on target vowels from a subset of data (nouns referring to special toys used in the lab to elicit corner vowels) may not tell the whole story about the purposes of IDS.

5. Acknowledgments

We thank the families who participated in the project.

6. References

[1] Fernald, A., Taeschner, T., Dunn, J., Papousek, M., de Boysson-Bardies, B., & Fukui, I. (1989). A cross-language study of prosodic modifications in mothers' and fathers' speech to preverbal infants. *J Child Lang*, *16*(3), 477-501.

[2] Katz, G. S., Cohn, J. F., & Moore, C. A. (1996). A combination of vocal f0 dynamic and summary features discriminates between three pragmatic categories of infant-directed speech. *Child Development*, *67*(1), 205-217.

[3] Cameron-Faulkner, T., Lieven, E., & Tomasello, M. (2003). A construction based analysis of child directed speech. *Cognitive science*, *27*(6), 843-873.

[4] Werker, J. F., Pons, F., Dietrich, C., Kajikawa, S., Fais, L., & Amano, S. (2007). Infant-directed speech supports phonetic category learning in English and Japanese. *Cognition*, *103*(1), 147-162.

[5] Uther, M., Knoll, M. A., & Burnham, D. (2007). Do you speak E-NG-LI-SH? A comparison of foreigner-and infant-directed speech. *Speech communication*, *49*(1), 2-7.

[6] Burnham, D., Kitamura, C., & Vollmer-Conna, U. (2002). What's new, pussycat? On talking to babies and animals. *Science*, *296*(5572), 1435-1435.

[7] Xu, N., Burnham, D., Kitamura, C. & Vollmer-Conna, U. (2015). Vowel hyperarticulation in parrot-, dog- and infant-directed speech. *Anthrozoos 26*, 3, p. 373-380

[8] Fernald, A., & Simon, T. (1984). Expanded intonation contours in mothers' speech to newborns. *Dev Psy*, *20*(1), 104.

[9] Papoušek, M., Bornstein, M. H., Nuzzo, C., Papoušek, H., & Symmes, D. (1990). Infant responses to prototypical melodic contours in parental speech. *Infant behavior and development*, *13*(4), 539-545.

[10] Stern, D. N., Spieker, S., & MacKain, K. (1982). Intonation contours as signals in maternal speech to prelinguistic infants. *Dev Psy*, *18*(5), 727.

[11] Fernald, A. (1989). Intonation and communicative intent in mothers' speech to infants: is the melody the message?. *Child development*, 1497-1510.

[12] Fernald, A. (1992). 13. Meaningful melodies in mothers' speech to infants. *Nonverbal vocal communication: Comparative and developmental approaches*, *262*.

[13] Werker, J. F., & McLeod, P. J. (1989). Infant preference for both male and female infant-directed talk: A developmental study of attentional and affective responsiveness. *Canadian Journal of Psychology/Revue canadienne de psychologie, 43*(2), 230–246.

[14] Fernald, A., & Mazzie, C. (1991). Prosody and focus in speech to infants and adults. *Dev Psy*, *27*(2), 209.

[15] Gleitman, L. R., Gleitman, H., Landau, B., & Wanner, E. (1988). Where learning begins: Initial representations for language learning. *Language: Psychological and biological aspects*, 150-193.

[16] Hirsh-Pasek, K., Nelson, D. G. K., Jusczyk, P. W., Cassidy, K. W., Druss, B., & Kennedy, L. (1987). Clauses are perceptual units for young infants. *Cognition*, *26*(3), 269-286.

[17] Nelson, D. G. K., Hirsh-Pasek, K., Jusczyk, P. W., & Cassidy, K. W. (1989). How the prosodic cues in motherese might assist language learning. *J Child Lang*, *16*(1), 55-68.

[18] Adriaans, F., & Swingley, D. (2017). Prosodic exaggeration within infant-directed speech: Consequences for vowel learnability. *JASA*, *141*(5), 3070-3078.

[19] Englund, K. T., & Behne, D. M. (2005). Infant directed speech in natural interaction—Norwegian vowel quantity and quality. *Journal of psycholinguistic research*, *34*(3), 259-280.

[20] Hartman, K. M., Ratner, N. B., & Newman, R. S. (2017). Infant-directed speech (IDS) vowel clarity and child language outcomes. *J Child Lang*, 44(5), 1140-1162.

[21] Zangl, R., & Mills, D. L. (2007). Increased Brain Activity to Infant-Directed Speech in 6-and 13-Month-Old Infants. *Infancy, 11*(1), 31-62. https://doi.org/10.1207/s15327078in1101_2

[22] Soley, G., & Sebastian-Galles, N. (2020). Infants' expectations about the recipients of infant-directed and adult-directed speech. *Cognition*, 198, 104214.

[23] Song, J. Y., Demuth, K., & Morgan, J. (2010). Effects of the acoustic properties of infant-directed speech on infant word recognition. *JASA*, *128*(1), 389-400.

[24] Weiyi Ma, Roberta Michnick Golinkoff, Derek M. Houston & Kathy Hirsh-Pasek (2011) Word Learning in Infant- and Adult-Directed Speech. *Lang Learn Dev*, 7:3, 185-201,

[25] Buchan, H., & Jones, C. (2014). Phonological reduction in maternal speech in northern Australian English: change over time. *J Child Lang*, *41*(4), 725-755.

[26] Liu, H. M., Tsao, F. M., & Kuhl, P. K. (2009). Age-related changes in acoustic modifications of Mandarin maternal speech to preverbal infants and five-year-old children: a longitudinal study. *J Child Lang*, *36*(4), 909-922.

[27] Kitamura, C., & Burnham, D. (2003). Pitch and communicative intent in mother's speech: Adjustments for age and sex in the first year. *Infancy, 4*(1), 85–110.

[28] Ratner, N. (1984). Patterns of vowel modification in mother–child speech. *J Child Lang*, 11(3), 557-578.

[29] Kidd, E., & Garcia, R. (2022). How diverse is child language acquisition research? *First Language*.

[30] Bohn, O. S. (2013). Acoustic characteristics of Danish infant directed speech *Proc. Mtgs. Acoust*. 19, 060055

[31] Sarvasy, H., Elvin, J., Li, W., & Escudero, P. (2019). An acoustic analysis of Nungon vowels in child-versus adult-directed speech. In *Proceedings of the ICPhS Melbourne* (pp. 3155-3159).

[32] Kitamura, C., & Thanavishuth, C. & Burnham, D., & Luksaneeyanawin, S. (2001). Universality and specificity in infant-directed speech: Pitch modifications as a function of infant age and sex in a tonal and non-tonal language. *Infant Behavior and Development*, 24. 372-392.

[33] Igarashi, Y., Nishikawa, K. Y., Tanaka, K., & Mazuka, R. (2013). Phonological theory informs the analysis of intonational exaggeration in Japanese infant-directed speech. *JASA*, *134*(2), 1283-1294.

[34] Bernstein Ratner, N., & Pye, C. (1984). Higher pitch in BT is not universal: Acoustic evidence from Quiche Mayan. *J Child Lang*, *11*(3), 515-522.

[35] Pye, C. (1986). Quiché Mayan speech to children. *J Child Lang*, *13*(1), 85-100.

[36] Bundgaard-Nielsen, R. L., & O'Shannessy, C. (2021). When more is more: The mixed language Light Warlpiri amalgamates source language phonologies to form a near-maximal inventory. *JPhon*, *85*, 101037.

[37] Butcher, A. R. (1994). On the phonetics of small vowel systems: evidence from Australian languages. In *Proceedings of the 5th SST* (Vol. 1, pp. 28-33). Canberra: ASSTA.

[38] Laughren, M. (1984). Warlpiri baby talk. *Australian Journal of Linguistics*, *4*(1), 73-88.

[39] Jones, C., & Meakins, F. (2013). The phonological forms and perceived functions of janyarrp, the Gurindji 'baby talk'register. *Lingua*, *134*, 170-193.

[40] Bundgaard-Nielsen, R. L., Baker, B. J., Kroos, C. H., Harvey, M., & Best, C. T. (2015). Discrimination of multiple coronal stop contrasts in Wubuy (Australia): A natural referent consonant account. *PLoS One*, *10*(12), e0142054.

[41] McLeod, S., & Crowe, K. (2018). Children's consonant acquisition in 27 languages: A cross-linguistic review. *American journal of speech-language pathology*, *27*(4), 1546-1571.

[42] Butcher, A., & Anderson, V. (2008). The vowels of Australian aboriginal english. In *Ninth International Speech Communication Association*.

[43] Polka, L., Masapollo, M., & Ménard, L. (2021). Setting the stage for speech production: Infants prefer listening to speech sounds with infant vocal resonances. *JSLHR*.

Apical stops in Arabana: lenition and undershoot

Mark Harvey[1], *Juqiang Chen*[2], *Michael Carne*[3], *Rikke Bundgaard-Nielsen*[1], *Clara Stockigt*[4], *Jane Simpson*[3], *Sydney Strangways*

[1]University of Newcastle, [2]Shanghai Jiao Tong University, [3]Australian National University, [4]University of Adelaide

mark.harvey@newcastle.edu.au, juqiang.c@stju.edu.cn, michael.carne@anu.edu.au, clara.stockigt@adelaide.edu.au, jane.simpson@anu.edu.au

Abstract

Commonly in Australian languages, the apical stops /t, ʈ/ show undershoot in manner, involving rhotic realizations: [r, ɾ, ɽ, ɻ], and in place involving the sublaminal target in retroflexion resulting in [t] realizations of /ʈ/. In Arabana, the apical stops rarely undershoot to rhotics, but undershoot realizations of /ʈ/ as [t] are common. Lenition theories posit that undershoot correlates with reduced duration. Lenition theories differ as to whether they posit undershoot or duration as the defining characteristic of lenition. Arabana expands the theory testing space as the manner undershoot is associated with reduced duration, but the place undershoot is not.

Index Terms: apicals, Australian languages, lenition, rhotics, stops, undershoot

1 Introduction

Arabana is a highly endangered language of northern South Australia. It is analyzed as having a contrast between an alveolar stop /t/ and a retroflex stop /ʈ/ within the apical category (see Table 1 for the full phonemic inventory) [1]. Lenition of intervocalic stops is a widespread phenomenon in Australian languages, and the apical stops commonly lenite to taps [2], [3]. In addition to the two apical stops, Arabana is also analysed as having three rhotics: an alveolar tap /ɾ/, an alveolar trill /r/, and a retroflex approximant /ɻ/.

	lab	den	alv	rfx	pal	vel
stop	p	t̪	t	ʈ	c	k
nasal	m	n̪	n	ɳ	ɲ	ŋ
lateral		l̪	l	ɭ	ʎ	
trill			r			
tap			ɾ			
approximant	w			ɻ	j	

Table 1: *Phonemic inventory of Arabana in IPA adapted from* [1]. *Apical stops and rhotics in grey. There are four phonemic vowels*: /i, u, a, a:/.

Apicals only appear in word-medial positions in Arabana [1]. The five-way contrast in stops and rhotics, /t/ vs /ʈ/ vs /r/ vs /ɾ/ vs /ɻ/, is found only in post-tonic, intervocalic position, i.e. ˈ#(C)V_V. Prosodic prominence is word-initial in Arabana and the tonic vowel is always the first vowel. In other intervocalic positions, there is a four-way contrast: /t/ vs /ʈ/ vs /r/ vs /ɻ/. Oppositions between apical stops and rhotics in consonant clusters are highly restricted and we do not consider data from clusters here.

1.1 Aims

This paper has two aims. The first aim is to quantify the extent of manner lenition, where apical stops are realized as taps, in Arabana. The second aim is to examine the overall patterns of gestural undershoot in the production of apical stops in Arabana and consider their implications for general theories of lenition.

If the common Australian patterns of lenition operate in Arabana, then the hypothesis is that there should be substantial overlap in realization between /t/ and /ɾ/ (see discussion in [4]). Given that lenition also commonly involves approximant realizations, there could also be overlap between /ʈ/ and /ɻ/, and [1] notes overlap in the realization of the three rhotics. We report quantitative data on the production of apical stops and rhotics in inter-vocalic positions in Arabana and examine whether these hypotheses are supported.

Gestural undershoot has played an important role in the debates over general theories of lenition, and in particular the question of whether all phenomena termed ‘lenition’ constitute a coherent class or not [5]. Some analyses propose that there is a coherent class of lenition processes with the defining criterion that processes involving gestural undershoot are lenition processes [6]. Other analyses propose that the defining criterion for lenition is variation in duration, which is principally conditioned by prosody, such that stronger prosodic positions favor longer, fortis realizations and weaker positions favor shorter, lenis realizations [7], [8]. Both analyses propose a relation between shorter duration and gestural undershoot.

General theories of lenition consider undershoot only from the perspective of manner of articulation. However, undershoot is also a factor from the perspective of place of articulation for apicals. In articulatory terms, retroflexes are more complex than alveolars. Both alveolars and retroflexes require apical contact, but retroflexes in Australian languages also consistently involve a sublaminal gesture [9]. If the sublaminal gesture is not attained, then an alveolar articulation results.

The Arandic and Western Desert languages associated with the areas to the north and west of Arabana are reported to have an alveolar versus retroflex opposition. However, in these languages,there is considerable overlap in the realizations of the two phonemes [10], [11]. We examine the distribution of alveolar versus retroflex phonetic realizations for apical stops in Arabana, and consider the role of gestural undershoot in this distribution.

2 Methods

2.1 Datasets

The analysis reported here is based on two datasets. In both datasets, the stimuli were real words. Dataset 1 consists of recordings with a first-language speaker of Arabana, Mr Sydney Strangways who was born in 1932, made in two fieldtrips in July 2019 and September 2020. In each fieldtrip, Author 5 initially reviewed a draft runsheet of headwords (extracted from [12]) with Mr Strangways. A final list of target headwords was then created. Each headword was assigned a unique identifier and a visual stimulus in MS Powerpoint was created. The headwords were randomized for five separate recording sessions made on consecutive days. Recordings were made in quiet locations with a Zoom H5 recorder using its internal microphone. To yield sufficient tokens for our analyses each elicitation prompted six to eight tokens (i.e. 30-40 tokens across sessions per type/condition).

Dataset 2 consists of a less structured recording with another first-language speaker of Arabana, Mr Laurie Stuart, now deceased, who was born in 1913. These audio materials were originally recorded to accompany a set of Arabana teaching materials already in print [13]. In order to provide comparison data with Dataset 1, we extracted target words with the same structure as were recorded in Dataset 1

Datasets 1 and 2 consisted of 57 and 60 distinct headwords respectively. The datasets included the full range of flanking vowels However, as the stimuli were real words, it was not possible to quantitatively balance the sets of flanking vowels. The three vowels /a, i, u/ do not have an equal distribution in the Arabana lexicon. In a vocabulary of 2,142 headwords, the vowel distribution is as follows: /a/ 4214 (57%), /i/ 1732 (24%), /u/ 1419 (19%) [12].

We distinguish three prosodic positions for word-medial inter- vocalic apicals in Arabana.

#(C +) V + Onset 1 + V + Onset 2 + V + Onset 3 + V

As illustrated, Onset 1 is the post-tonic onset and Onsets 2 and 3 are progressively later in the word. Both datasets sampled all three prosodic positions, but sampling of the positions was not quantitatively balanced because real word stimuli were required.

2.2 Annotation procedure and word selection

All target segments were manually segmented and transcribed by two phonetically trained annotators in *Praat 6.0.43* with the same spectrogram settings (Frequency range = 0-8 kHz; Dynamic range = 40.0 dB, window length 0.005; mean intensity (db) overlayed). The annotators were provided with orthographic transcriptions of the target words, but the target segments were masked.

The phonetic transcription was based a set of acoustic criteria. Place opposition between alveolar and retroflex articulation was distinguished using two well-known acoustic correlates of retroflex articulation: lowering of F3 ([14], [15]); and/or convergence of F3-F2 on the proceeding vocalic segment. Taps and trills were identified spectrographically by the following criteria: (1) a reduction in the amplitude (dB) of waveform relative to the spectral envelope; (2) a corresponding drop in mean spectral energy (in dB); (3) loss or attenuation of formant visible in the spectrogram between 500–5000 Hz. Stops were distinguished from taps by clear evidence of significant closure duration, with or without release.

In Dataset 1, from a pool of 2791 tokens, annotators agreed on 2583 transcriptions (92.5%). In Dataset 2, from a pool of 351 tokens, annotators agreed on 335 transcriptions (95%). These agreed tokens are the basis for the following analysis.

3 Results & Analysis

3.1 Distribution of phonetic stop and rhotic realizations against the phonological manner opposition /stop/ vs /rhotic/

Table 2 shows the phonetic and phonological distribution of stops and rhotics in Dataset 1. There was a high degree of consistency between the blind phonetic transcription of the two annotators and the phonological analysis of the target segments in Arabana. Of the 462 target segments transcribed phonetically as stops [t, ʈ], 458 (99%) were realizations of phonological stops /t, ʈ/. Of 548 phonological stop tokens /t, ʈ/, 458 (84%) were realized by phonetic stops [t, ʈ]. There was considerable variation in the remaining 16% of phonological stop realizations: [ɾ, ɹ, ɽ, ɻ]. It may be noted that the set of phonetic rhotic realizations for both phonological stops and rhotics includes the alveolar approximant [ɹ], which does not correspond to an independent phonological category.

Phonological category		Phonetic realization		Token No	%
Stop	/t, ʈ/	Stop	[t, ʈ]	458	83.6
		Rhotic	[ɾ, ɹ, ɽ, ɻ]	90	16.4
Rhotic	/r, ɾ, ɻ/	Stop	[t]	4	0.3
		Rhotic	[r, ɾ, ɹ, ɻ]	1,444	99.7

Table 2: *Distribution of stops and rhotics in Dataset 1*

Table 3 shows the phonetic and phonological distribution of stops and rhotics in Dataset 2.

Phonological category		Phonetic realization		Token No.	%
Stop	/t, ʈ/	Stop	[t, ʈ]	87	91.6
		Rhotic	[ɾ, ɻ]	8	8.4
Rhotic	/r, ɾ, ɻ/	Stop	[t]	2	1.5
		Rhotic	[r, ɾ, ɹ, ɻ]	132	98.5

Table 3: *Distribution of stops and rhotics in Dataset 2*

The same observation holds for Dataset 2 as for Dataset 1, with 87 of 89 (98%) of phonetic stop tokens being realizations of phonological stops, and 87 of 95 (92%) of phonological stop tokens being realized by phonetic stops.

3.2 Distribution of rhotic realizations of phonological stops by prosodic position

As discussed, we distinguish three prosodic positions for word-medial inter- vocalic apicals in Arabana. Onset 1 is the post-tonic onset and Onsets 2 and 3 are progressively later in the word. There was not sufficient data in Dataset 2 to evaluate

the correlations between prosody and rhotic realizations and consequently we report only on Dataset 1 here. Table 4 sets out the distribution of stop and rhotic realizations for stop phonemes by prosodic position in Dataset 1.

	Stop realizations		Rhotic realizations	
	No	Percentage	No	Percentage
Onset 1	324	93.1	24	6.9
Onset 2	125	70.6	52	29.4
Onset 3	9	39.1	14	60.9

Table 4. *Distribution of stop and rhotic realizations by prosodic position in Dataset 1*

Overall, rhotic realizations are minority realizations (16%) for stop phonemes in Dataset 1. As Table 4 shows this minority realization pattern is not evenly distributed. Rather, the weaker the prosodic position, the more likely it is that an apical stop phoneme is produced phonetically as a rhotic.

3.3 Duration of phonetic stop realizations by prosodic category

There was a significant durational contrast between phonetic stop realizations in Onset 1 and Onset 2 positions, as set out in Figure 2. There was not sufficient data to evaluate Onset 3.

Figure 2: *Phonetic stop realizations in Onset 1 & 2*

The duration measures were selected as the dependent variable and fitted with a general linear model in *R* [16]. Onset conditions and speakers (as in different datasets) were fixed factors. There were significant main effects of onset conditions ($F = 119.98$, $df = 1$, $p < 0.001$) and speakers ($F = 81.86$, $df = 1$, $p < .001$) and a significant onset × speaker interaction ($F = 5.24$, $df = 1$, $p = 0.02$).

To further examine the phonetic categories effect, we ran pairwise multiple comparisons with Tukey adjustments for the phonetic category differences. Stops had longer durations than taps for both dataset (dataset 1: $\beta = 0.02$, $SE = 0.002$, $t(505) = 10.85$, $p < 0.001$; and dataset 2: $\beta = 0.011$, $SE = 0.004$, $t(505) = 2.76$, $p = 0.03$).

3.4 Distribution of alveolar vs retroflex phonetic stop realizations as against the phonological alveolar vs retroflex stop categories

Table 5 sets out the distribution of alveolar and retroflex stop realizations for Dataset 1 and Dataset 2.

	[t]		[ʈ]	
	No	Percentage	No	Percentage
Dataset 1				
/t/	259	100		
/ʈ/	150	75.4	49	24.6
Dataset 2				
/t/	16	100		
/ʈ/	56	78.9	15	21.1

Table 5: *Alveolar vs retroflex stop realizations of alveolar vs retroflex stop phonemes*

In both datasets, retroflex stop realizations are a minority phenomenon, 49 of 458 (11%) of total stop realizations in Dataset 1 and 15 of 87 (17%) of total stop realizations in Dataset 2. In both datasets, [ʈ] realizations are found only with the /ʈ/ segment and even with the /ʈ/ segment, they are a minority phenomenon, constituting 20–25% of /ʈ/ realizations.

There was variation between the two datasets in type–token relations for [ʈ] tokens. In Dataset 1, the 49 [ʈ] realizations were found with only four word types: *kaʈijapu* 'head', *kuʈa-ɳʈa* 'lie-PRES', *ŋuʈi-ɳʈa* 'halt-PRES', *ɲampaʈa-ɳʈa* 'cover-PRES'. In Dataset 2, the 15 [ʈ] tokens were found with 8 word types.

There was no significant difference in duration between the alveolar stop and retroflex stop phonetic realizations of the phonological retroflex stop.

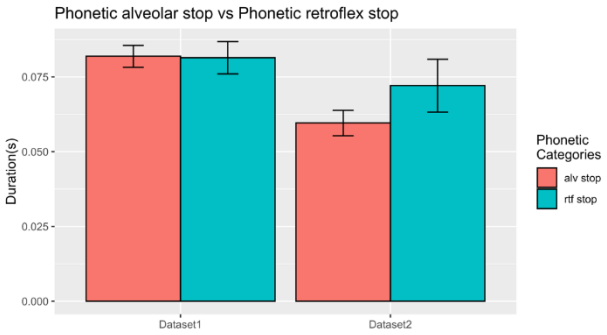


Figure 3: *Durations of phonetic alveolar stop and phonetic retroflex stop of the phonological retroflex stop*

The duration measures were selected as the dependent variable and fitted with a general linear model in *R* [16]. Manner of articulation and speakers (as in different datasets) were fixed factors. There is a significant main effect of speakers ($F = 48.29$, $df = 1$, $p < .001$) but no significant main effect of manner of articulation nor manner of articulation × speaker interaction.

4 Discussion

The two datasets show close agreement. In both datasets, there is a strong association between the phonetic and phonological stop categories, as summarized in Table 6.

	Dataset 1	Dataset 2
[t, ʈ] as realizations of /t, ʈ/	99%	98%
/t, ʈ/ realized by [t, ʈ]	84%	92%

Table 6*: Phonetic and phonological stops*

In both datasets, phonetic stop realizations in the post-tonic Onset 1 position are longer than phonetic stop realizations in the later Onset 2 position. In Dataset 1, where there was

sufficient data to evaluate the correlations between prosody and rhotic realizations, 66/90 (73%) of rhotic realizations of phonological stops occur in Onset 2 and 3 positions. Overall, there is an association between stronger prosodic positions, greater probability of stop realizations, and longer duration of stop realizations. Conversely, there is an association between weaker prosodic positions, greater probability of rhotic realizations, and shorter duration of stop realizations.

Unlike the manner opposition between stop and rhotic, the alveolar versus retroflex place opposition did not show a clear match between phonetic and phonological categories. In both datasets, the retroflex stop [ʈ] was found only as a realization of /ʈ/. However, [ʈ] constituted only 20–25% of stop realizations of /ʈ/, with the bulk of stop realizations of /ʈ/ being [t], as summarized in Table 7.

	Dataset 1	Dataset 2
/ʈ/ realized by [ʈ]	75%	79%
/t/ realized by [ʈ]	0%	0%

Table 7*: Realizations of apical stops by place*

There is one divergence between Datasets 1 and 2 relating to type–token relations for /ʈ/. In Dataset 1, the 49 [ʈ] tokens related to only four word types whereas in Dataset 2, the 15 [ʈ] tokens related to eight word types. This suggests that there may be significant inter-speaker variation in Arabana in the distribution of [ʈ] realizations across type–token relations. Further, there may be significant inter-speaker variation in the comparative frequencies of [t] and [ʈ] as realizations of /ʈ/.

5 Conclusions

Theories of lenition posit a correlation between shorter duration and gestural undershoot but differ as to whether variation in duration or undershoot is the defining characteristic of lenition. Apical stops in Arabana show variation in both manner and place of articulation. Both types of variation involve undershoot. In terms of manner, the rhotic realizations of phonological apical stops can be modelled as a failure to attain the complete closure required for a phonetic stop realization. In terms of place, the alveolar realizations of the phonological retroflex stops can be modelled as a failure to attain the additional sublaminal gesture which distinguishes retroflexes from alveolars.

If undershoot is the defining characteristic of lenition, then both the rhotic realizations of the apical stops and the alveolar realizations of retroflex stops would class together as examples of lenition. However, it is only with the manner variations that undershoot correlates with shorter duration. For the place variations, alveolar realizations are not shorter than retroflex realizations. If variation in duration is the defining characteristic of lenition, then the manner variations are examples of lenition, but the place variations are not.

A full evaluation of undershoot vs. variation in duration as a basis for a category is beyond the scope of this paper. However, we note that analyses which take variation in duration as the criterion for category membership can encompass a wide range of phenomena extending beyond the traditional category of lenition to also include the traditional counterpart category of fortition [5], [7], [8]. By contrast, analyses taking undershoot as the criterion for category membership do not currently appear to encompass phenomena beyond undershoot. Given that analyses based in variation in duration appear to encompass a wider range of phenomena, they are favored, and we analyze the manner variations found with apical stops as lenition, but not the place variations.

Given this definition of lenition, apical stops in our Arabana data depart from typical patterns reported for Australian languages in showing only minimal lenition. The extent to which the patterns in our data are typical of Arabana more generally is a topic for further research. There are field recordings involving a number of other first language speakers of Arabana. These materials are not experimental materials and are generally similar in structure to Dataset 2.

These materials would also support further investigation into the nature of variation in the realization of retroflexion in Arabana. Our Arabana data on the apical place opposition was similar to that reported for the Arandic and Western Desert languages to the north and west in that there was no consistent match between phonetic and phonological categories [10], [11]. However, our Arabana data differed from that reported for the Arandic and Western Desert languages in that the inconsistency in matching was unidirectional. In Arabana, /ʈ/ commonly has [t] realizations, but /t/ does not have [ʈ] realizations. By contrast in the Arandic and Western Desert languages while /ʈ/ may have [t] realizations, /t/ may also have [ʈ] realizations.

A more general topic for further research is the extent to which lenition does and does not operate among Australian languages. Research on lenition in Australian languages has focused on contexts which satisfy the general phonological requirements for lenition and where lenition is commonly attested. There has been little investigation as to whether there are systematic correlates to the presence of lenition as a common phenomenon in these contexts. Equally, there has been very little investigation of situations where the general phonological requirements for lenition are satisfied but lenition is rarely attested. A better understanding of the comparative quantitative distribution of lenition from a cross-linguistic perspective is central to advancing theories of lenition.

6 Acknowledgments

We acknowledge with respect and gratitude the work of †Laurie Stuart and †Luise Hercus in recording Arabana. We thank Greg Wilson for access to his recordings of Arabana, and for discussion of Arabana language. We acknowledge that the Arabana language is the property of Arabana people

This research was supported by two grants from the ARC Centre of Excellence for the Dynamics of Language: (i) LDG992020 “Local vs Long-distance processes: Nasals and Nasalization in Arabana”; (ii) CE140100041 “Metrical Prominence and Pre-stopping in Arabana”. It was also supported by ARC DP190100646 “1 potato, 2 wotatoes, 3 otatoes: Lexical access in Australian languages”.

7 References

[1] L. Hercus, *A grammar of the Arabana-Wangkangurru language Lake Eyre Basin, South Australia*, vol. 128. Canberra: Pacific Linguistics, 1994.

[2] B. Baker, “Word structure in Australian languages,” in *The languages and linguistics of Australia: a comprehensive guide*, H. Koch and R. Nordlinger, Eds. Berlin & Boston: Walter de Gruyter, 2014, pp. 139–213.

[3] R. M. W. Dixon, *Australian languages: their nature and development*. Cambridge: Cambridge University Press, 2002.

[4] 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.
[5] J. Katz, "Lenition, perception and neutralisation," *Phonology*, vol. 33, pp. 43–85, 2016.
[6] L. Bauer, "Lenition revisited," *J. Linguist.*, vol. 44, no. 3, pp. 605–624, 2008.
[7] J. Katz, "Intervocalic lenition is not phonological: evidence from Campidanese Sardinian," *Phonology*, vol. 38, no. 4, pp. 651–692, Nov. 2021, doi: 10.1017/S095267572100035X.
[8] U. C. Priva and E. Gleason, "The causal structure of lenition: A case for the causal precedence of durational shortening," *Language*, vol. 96, no. 2, pp. 413–448, 2020, doi: 10.1353/lan.2020.0025.
[9] A. Butcher, "The phonetics of neutralisation: the case of Australian coronals," in *Studies in general and English phonetics*, J. Windson, Ed. London & New York: Routledge, 1995, pp. 10–38.
[10] M. Tabain and R. Beare, "An articulatory study of the alveolar versus retroflex contrast in pre- and post-stress position in Arrernte," *J. Phon.*, vol. 78, p. 100952, Jan. 2020, doi: 10.1016/j.wocn.2019.100952.
[11] M. Tabain, A. Butcher, G. Breen, and R. Beare, "A formant study of the alveolar versus retroflex contrast in three Central Australian languages: Stop, nasal, and lateral manners of articulation," *J. Acoust. Soc. Am.*, vol. 147, no. 4, pp. 2745–2765, Apr. 2020, doi: 10.1121/10.0001012.
[12] L. Hercus, "Arabana dictionary."
[13] G. Wilson, *Audio-recordings of Laurie Stuart, Jean Wood, Millie Warren, and Pauline Thompson speaking/pronouncing 5000 Arabana words, phrases and interactive sentences.* Lodged with the Arabana Corporation, 2005.
[14] S. Hamann, *The Phonetics and Phonology of Retroflexes*. Utrecht: LOT Publishing, 2003.
[15] M. Lindau, "The story of /r/," in *Phonetic Linguistics: Essays in honor of Peter Ladefoged*, Cambridge, MA: Academic Press, 1985, pp. 157–168.
[16] R Core Team, "R: A language and environment for statistical computing." Vienna, 2018. [Online]. Available: https://www.R-project.org/

L2-Mandarin regional accent variability during lexical tone word training facilitates naive English listeners' tone categorization and discrimination

Yanping Li[1], *Catherine T. Best*[1,2], *Michael D. Tyler*[1,3], *Denis Burnham*[1]

[1] The MARCS Institute, Western Sydney University, Australia
[2] Haskins Laboratories, New Haven, U.S.A.
[3] School of Psychology, Western Sydney University, Australia

yanping.li@westernsydney.edu.au, c.best@westernsydney.edu.au
m.tyler@westernsydney.edu.au, denis.burnham@westernsydney.edu.au

Abstract

Mandarin-naïve English listeners have difficulties categorizing the four lexical tones that distinguish word meanings in Mandarin. This study investigated how L2-Mandarin regional accent variability in training on minimal-tone-contrast words affected tone perception. Prior to training, although listeners accurately categorized and discriminated rising and dipping tones, they confused falling and level tone significantly more than the other tone contrasts. After training, learners in the accent variability (experimental) condition showed improved categorization and discrimination of falling and level tones; constant-accent (control condition) learners did not. The results supported the hypothesis that accent variability during lexical tone word training facilitates tone categorization.

Index Terms: high vs low variability training; regional accents; lexical tone discrimination; tone contour categorization

1. Introduction

There are four lexical tones in Mandarin [1] that differ in their fundamental frequency (f_0) patterns with f_0 height and f_0 contour as the primary acoustic parameters[2]: T1 high level contour, T2 high rising, T3 low dipping and T4 high falling. Tones are used to distinguish word meanings (e.g., for the consonant-vowel [CV] syllable /ma/, level = *mother*, rising = *hemp*, dipping = *horse*, and falling = *to scold*). Listeners of languages that lack tones at the pre-lexical level, e.g., English or French listeners, perceptually assimilate tones to their native intonational categories [3], [4]. For example, the rising and falling tones are assimilated to English question (yes-no) versus statement intonations, respectively, due to phonetic similarities. Thus, non-tone language listeners are not entirely "deaf" to tones. When categorizing tones according to perceived pitch/contour, without referring to native intonational categories, their tone perception is psychophysically based [5], reflected in varying performance across tones [6]–[8]. Specifically, while listeners can categorize the pitch/contour of tone contrasts that display acoustic dissimilarities, such as T1 level vs. T3 dipping, T2 rising vs. T4 falling and T3 dipping vs. T4 falling, they have difficulties with those that show acoustic similarities, e.g., T1 level vs. T4 falling, T1 level vs. T2 rising, and T2 rising vs. T3 dipping [9].

Despite initial difficulties in categorizing Mandarin tones, naive English listeners can show improvements after high-variability perceptual training with multiple talkers. In [10], tone categorization improved 21% from pre- to post-test after high-variability training, which was maintained 6 months later. They also showed generalization to novel stimuli from one of the talkers used in training and to a novel talker. Based on this result, high-variability perceptual training has been adapted in other second language (L2) suprasegmental training studies (e.g., [11], [12]).

One limitation of perceptual training for improving L2 acquisition is that the training focuses on tone categorization rather than their lexical relevance, i.e., tone word identification. Meaningful words, rather than isolated (supra)segments, are employed in conversations, which creates coordinated patterns of activity in sensory and higher level cognitive functions [13]. In tasks involving simple tone categorization, even high-variability perceptual training still fails to relate tonal form to lexical meaning. Training with words that contrast only in their lexical tones can address this shortcoming. In a picture-to-word L2 training paradigm, English learners trained on Spanish words produced by multiple talkers showed significantly better accuracy in identifying target words than those trained with just a single talker [14]. Thus, high talker variability can facilitate L2 word learning in a non-tone language. One study of potential relevance to talker variability effects on minimal-tone-contrast word learning trained English learners to identify pseudowords recorded by English speakers, in which the pitch contours had been resynthesized. Participants with high-variability (multiple talkers) training achieved higher post-training accuracy than those with low-variability (one talker) training [15].

If high talker variability word training helps English listeners identify minimal-contrast words differing only in tone, it should in turn promote tone categorization and discrimination for each of the six tone contrasts (T1 vs. T2, T1 vs. T3, T1 vs. T4, T2 vs. T3, T2 vs. T4, T3 vs. T4), which was not investigated in prior studies either on tone perception [6]–[9] or on Mandarin minimal-tone-contrast word training [15], [16]. If English learners can identify words differing only in tone after training, implying that they have established tone categories, they should be able to use those categories to sort and discriminate tones at the pre-lexical level. We tested this hypothesis by examining tone categorization and discrimination after word training.

While our training used high talker variability, we manipulated degree of accent variability, which has not been examined before. Accent variability in tones is triggered by similarities and dissimilarities between the tone systems of Chinese regional dialects and standard Beijing Mandarin [17], [18] which regional speakers learn as an L2. For example, Yantai, Shanghai and Guangzhou speakers produce their L2 Mandarin dipping tone with shallower falling-rising contours than Beijing Mandarin productions (see [18] for acoustic details). This study investigates how L2-Mandarin regional

accent variability affects English listeners' L2 tone categorization and discrimination following word training. We posit that exposure to L2-Mandarin regional accent variability will facilitate word learning.

In our study, English participants in two conditions were trained on minimal-tone-contrast words with high talker variability. The control group heard only Beijing-accented stimuli, whereas the experimental group heard them spoken with Beijing and another two L2-Mandarin regional accents. Their tone contour categorization and discrimination in pre- and post-training conditions were estimated under the framework of Perceptual Assimilation Model (PAM, [19], [20]), which focuses on perceptual assimilations to native phoneme categories in cross-language speech perception by (naïve) L2 listeners and PAM has been extended to lexical tone assimilation [21], [22]. Prior to training, English listeners should perceptually assimilate tones as Non-Assimilable nonspeech pitch heights and contours, because there are no lexical tone categories in their phonological system. Nonspeech tone icons presented in [22] will be used for categorization. Naïve English listeners should assimilate each tone to several tone icons sharing height/contour similarities, which are learnable according to PAM-L2 [19]. Given this and four icons, although they are expected to be initially Non-Assimilable to phonological categories, they can nonetheless be either Categorized or Uncategorized to specific nonspeech icons. Their discrimination of the four Mandarin tones is expected to be good to very good. After minimal-tone-contrast word training, learners in both conditions should be able to abstract the tones as lexically relevant L2 phonological categories and their assimilation types are expected to shift to being Categorized or Uncategorized as L2 phonological components in speech with experimental listeners performing better than the control group due to effects of L2-Mandarin regional accent variability. Their tone discrimination varies based on tone assimilation types and training conditions. Specifically, listeners in the experimental group are more likely to assimilate each tone to a different tone contour category and their discrimination will be excellent. Listeners in the control group should easily Categorize tones with dissimilarities. On the contrary, they may show Uncategorized assimilation of tones with height/contour similarities, such as T1 level vs. T4 falling, resulting in Categorized-Uncategorized assimilation with good discrimination.

2. Experiment

2.1. Method

2.1.1. Participants

Mandarin-naïve English speakers (n = 48) were recruited online for this study, and randomly assigned to the single accent (control: n = 24, M_{age} = 24.5 years, SD = 5.8 years, 14 females) or multiple accent (experimental: n = 24, M_{age} = 25.5 years, SD = 5.1 years, 15 females) conditions. All were functional monolinguals [9] from English-speaking countries, primarily Australia. Prior to this study, none had experience with any tone languages, e.g., Mandarin, Vietnamese, Thai, Cantonese. Since musical training can facilitate tone perception [15], [23], none had more than 3 years of private lessons in any combination of instruments [15]. Their language and music backgrounds were self-reported through an online survey, which also determined that none had speech or hearing disorders. They received Prezzee eGift smart cards for their participation.

2.1.2. Stimuli

There were 16 Mandarin tone real words for pre- vs. post-categorization and discrimination tasks, which were generated from four CV syllables (/ga/, /ti/, /tu/, /pu/) with four tones. The three vowels /a/, /i/, /u/ were selected, because they are used in both Mandarin and English. Word targets and their characters were very regular based on [24] and they were produced by a native Beijing female talker (Age = 25.0 years) in a soundproof booth at the Speech Acquisition and Intelligent Technology Lab, Beijing Language and Culture University, Beijing, China. Apparatus and procedures for recordings were the same as those in [18]. Her productions were verified by four other native Beijing female listeners (M_{age} = 20.75 years, SD = 2.49 years) on a scale of 1 (not clear) to 7 (clear) and only the four tokens with the highest ratings for each word were retained, resulting in 64 (16 words × 4 tokens) stimulus items.

Another set of 16 Mandarin tone real words (four CV syllables, /ba/, /di/, /du/, /gu/ × 4 tones) were used in the Mandarin tone word training task. They were produced by native female talkers selected from [18], either 12 from Beijing (control: constant accent, talker-only variability) or four each from Beijing, Yantai, and Guangzhou (experimental: accent and talker variability). Both groups thus heard 12 talkers. As with the pre- and post-training stimuli, training word productions were verified by four other native female listeners of each dialect and final tokens were selected, resulting in 768 (12 speakers × 16 words × 4 tokens) stimuli in each training group. Word meanings for training were indicated by grey-scale pictures from [25], counterbalanced across participants, resulting in 16 pseudowords per participant.

2.1.3. Procedure

This study was conducted remotely with E-prime Go 1.0. Participants ran the perceptual tasks on their own Windows 10 laptops/desktops. To ensure data quality, they were tested through a ZOOM meeting with the experimenter, using wired (not bluetooth/wireless) earphones in a quiet room.

Learners completed Mandarin tone word training, pre- and post-training tone categorization and discrimination tasks, and post-training word verification and generalization tests (not reported here). The focus of this study was how L2-Mandarin regional accent variability affects the listeners' pre- to post-training tone perception, so the training procedure is described here briefly. They learned the 16 tone pseudowords in a picture-to-word paradigm, which were produced by 4 talkers in each training session (45 minutes), with the selected groups of talkers counterbalanced across the six training sessions that used quizzes with feedback [15]. Experimental group talkers were blocked by regional accent. Correspondingly, talkers for the control group were randomly assigned to subgroups of four with the *R* [26] *Sample* function.

A 1-minute tone familiarization was presented at the start of the pre-training perceptual test to acquaint listeners with the four tone contour icons. They then completed the tone discrimination and categorization tests in that order. As in studies investigating PAM-based predictions for consonants [27] and vowels [28], in the pre- and post-training perceptual tests, listeners first completed a categorical AXB discrimination test for each of the six tone contrasts, which were blocked with a Latin Square design across participants. In each trial, A and B were tokens of the contrasting tone categories, and the middle item (X) was the same tone as the first (A) or third (B) item; interstimulus interval (ISI) was 1 s. Listeners were asked to

click on the "1" or "3" displayed on the screen to indicate whether the X item matched category A or B. To avoid simple acoustic identity judgements, the X item was a different token of the same tone category as the matching A or B item. For each contrast, the four AXB trial types (AAB, ABB, BBA, BAA) occurred four times, and each of the four tokens per stimulus set occurred twice in each position (first, second, or third). There were 64 (4 syllables × 4 AXB trial types × 4 times) randomized trials for each of the six tone contrast blocks, resulting in a total of 384 (6 tone contrasts × 64 trials) trials.

Following a short break, participants then completed the tone categorization task using four tone icons displayed in the four quadrants of their screens. First came eight practice items (two tokens for each tone category) in random order, followed by the 64 (16 words × 4 tokens) trials of individual test items. The response time-out was 3.5 s. Participants were told to click a button in the centre of their screen to activate each trial, which was equidistant from the four tone contour icons, and one token of one of the 16 test pseudowords played out. Listeners then indicated which tone they had heard by clicking on one of the four tone icons. The tone icon positions were held constant for each participant and counterbalanced across participants. Listeners were instructed to hold the central activation button until they heard the stimulus. If they released it too soon, the trial ended automatically. Trials that ended automatically or lacked a response within 3.5 s (178 occurrences, or 2% of all trials) were repeated at the end of each block. Listeners were trained on the target words in consecutive six days with no more than two training sessions conducted in the same day. After completing the last training session, they completed the same tone discrimination and categorisation tests as during pre-training.

2.2. Results

2.2.1. Identification of Mandarin tones

Figure 1 shows mean percentage of choices for the four Mandarin tones in pre- and post-training tests by participants in the experimental (accent variability) and control (Beijing-only) word training conditions. Two criteria were used to determine tone assimilations [21] to the icons: (1) a given f_0 contour icon must be selected significantly more than chance level (25%), and (2) it must be chosen significantly more often than any other icons. For each group in each test phase (pre, post), separate one-sample *t*-tests against chance level 25% were

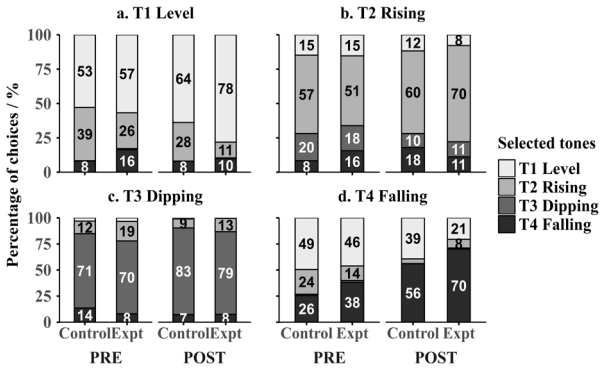


Figure 1: *Mean percentage of choices of the four tone icons for Mandarin tones in pre- and post-training tests by listeners in the experimental (accent variability) and control (Beijing-only) word training conditions.*

conducted for each tone to address criterion (1) using *R* with the *Student's t-Test* function. Multiple linear mixed-effects models were built for criterion (2) with the *lmer* function from package *lme4* [29]. Participants' percentage of choices for each icon was specified as the dependent variable. Training conditions, test phases, and the selected Mandarin tones were specified as fixed effects, and participants as a random effect. The Kenward-Roger approximation to the degrees of freedom was used to calculate the *p* values for the fixed-effects factors [30] and the *Anova* function from package *car* [31] was used to calculate *F*. Pairwise comparisons were conducted with *lsmeans* [32] in *R* whenever necessary to determine whether the percentage of choosing tone icon was greater than each of the other icons.

When responding to T1 level tone stimuli in the pre-training test, the listeners in both conditions split their choices between level (M_{expt} = 56.77%; $M_{control}$ = 52.86%) and rising (M_{expt} = 26.04%; $M_{control}$ = 38.80%) icons. Level icons were selected significantly above chance ($t(23) = 6.69$, $p < .001$) only in the experimental condition, and were selected significantly more often than the rising icons, Estimate = 30.73, $SE = 5.99$, $t(345) = 5.13$, $p < .0001$, indicating that these listeners Categorized T1 level tone before training. In the control condition, level and rising icons were both selected significantly above chance (T1: $t(23) = 5.44$, $p < .001$; T2: $t(23) = 2.77$, $p = .005$) and there was a non-significant difference between them (Estimate = 14.06, $SE = 5.99$, $t(345) = 2.34$, $p = .58$), indicating that T1 level tone stimuli were Uncategorized by the control participants. After training, for T1 stimuli only the level icon was selected significantly more than 25% by each group (M_{expt} = 78.12%, $t(23) = 9.49$, $p < .001$; $M_{control}$ = 63.08%, $t(23) = 5.37$, $p < .001$), suggesting that T1 level tone assimilation became Categorized for both groups. For responses to T2 rising and T3 dipping tone stimuli, listeners in both conditions selected rising and dipping icons, respectively, in both pre- and post-training tests above 50% which is significantly above chance, and each was chosen significantly more often than the other three icons, indicating that T2 rising and T3 dipping were correctly Categorized.

When responding to T4 falling tone stimuli, both groups split their choices between falling (M_{expt} = 38.02%; $M_{control}$ = 25.52%) and level (M_{expt} = 46.09%; $M_{control}$ = 49.47%) icons. Prior to training, participants in the experimental condition selected both falling ($t(23) = 2.51$, $p = .009$) and level ($t(23) = 4.25$, $p < .001$) icons significantly above chance and the difference of choices between them was non-significant (Estimate = 8.07, $SE = 6.42$, $t(345) = 1.25$, $p = 0.99$), indicating that T4 falling was Uncategorized. However, in the control condition only the level icon ($t(23) = 5.09$, $p < .001$) was selected significantly above chance, which was also selected significantly more often than the other three tone icons, indicating that the listeners Categorized T4 falling incorrectly to the level icon. After training, the experimental condition listeners categorized T4 falling tone as the falling icon. They selected only falling icons significantly more than chance (M = 70.05%; $t(23) = 5.98$, $p < .001$), which was also selected significantly more often than the other three tones, indicating that the listeners had correctly Categorized T4 as falling. While the control condition listeners improved by shifting part of their level-icon choices to falling icons, they still showed Uncategorized assimilation of T4 falling stimuli, because they chose both falling (M = 55.98%; $t(23) = 4.11$, $p < .001$) and level (M = 39.32%; $t(23) = 2.04$, $p = .02$) icons more often than chance, but with no significant difference between them (Estimate = -16.66, $SE = 6.42$, $t(345) = -2.59$, $p = .40$).

2.2.2. Discrimination of Mandarin tones

Figure 2 displays tone discrimination in both conditions for the pre- and post-training tests. To control for response bias, each participant's discrimination data were transformed to *A* scores [33]:

$$A = \begin{cases} \frac{3}{4} + \frac{H-F}{4} - F\,(1-H) & \text{if } F \le 0.5 \le H; \\ \frac{3}{4} + \frac{H-F}{4} - \frac{F}{4H} & \text{if } F \le H < 0.5; \\ \frac{3}{4} + \frac{H-F}{4} - \frac{1-H}{4(1-F)} & \text{if } 0.5 < F \le H \end{cases} \quad (1)$$

Where *F* is false alarm rate and *H* is hit rate (see [34] for details on Signal Detection Theory). Higher *A* values indicate better discrimination.

Multiple linear mixed-effects models were built on those values, with training conditions, test phases, and tone contrasts being specified as fixed effects and participants as a random effect. Calculation of *p* and *F* values was the same as in the tone categorization model. While there was no significant main effect of training conditions, the main effects of test phases, $F(1, 550) = 5.01$, $p = 0.02$, and tone contrasts, $F(5, 546) = 14.05$, $p < .001$, and the training conditions × test phases × tone contrasts interaction, $F(23, 528) = 3.67$, $p < .001$, were all significant. To tease the interaction apart, pairwise comparisons were conducted to assess improvement in tone discrimination between pre- and post-training tests for both groups.

Prior to training, while listeners in both groups were highly sensitive to tone differences, they showed significantly lower sensitivity to differences between T1 level vs. T4 falling ($M_{expt} = 0.91$ in *A* scores; $M_{control} = 0.90$) than between T1 level vs. T3 dipping ($M_{expt} = 0.97$, Estimate = 0.07, $SE = 0.02$, $t(528) = 4.24$, $p = .006$; $M_{control} = 0.98$, Estimate = 0.07, $SE = 0.02$, $t(528) = 4.61$, $p = .001$); and T2 rising vs. T3 dipping ($M_{expt} = 0.97$, Estimate = -0.06, $SE = 0.02$, $t(528) = -3.99$, $p = .02$; $M_{control} = 0.97$, Estimate = -0.07, $SE = 0.02$, $t(528) = -4.03$, $p = .01$); and T3 dipping vs. T4 falling ($M_{expt} = 0.97$, Estimate = -0.06, $SE = 0.02$, $t(528) = -3.85$, $p = .03$; $M_{control} = 0.96$, Estimate = -0.06, $SE = 0.02$, $t(528) = -3.80$, $p = .03$). *A* scores of tone pairs T1 level vs. T2 rising and T2 rising vs. T4 falling for both groups fell between those of tone pair T1 level vs. T4 falling and tone pairs involving T3 dipping. No significant differences between them were observed. While listeners in both groups showed no significant improvement of sensitivity to tone differences in each tone pair after training, differences among tone pairs were

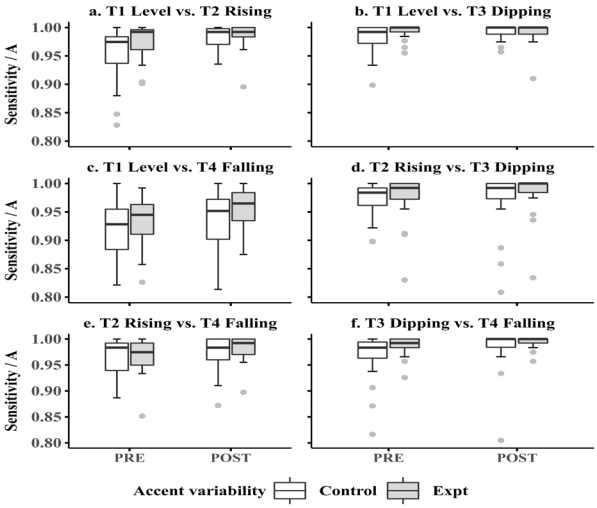


Figure 2: *AXB discrimination for the six tone contrasts in pre- and post-training tests by listeners in the experimental (accent variability) and control (Beijing-only) tone word training conditions. Outliers are displayed as grey points.*

no longer significant in the experimental condition, whereas listeners in the control condition continued to display less sensitivity to tone contrast T1 level vs. T4 falling ($M = 0.92$) than to T1 level vs. T3 dipping ($M = 0.97$; Estimate = 0.06, $SE = 0.02$, $t(528) = 3.55$, $p = .06$) and T3 dipping vs. T4 falling ($M = 0.98$; Estimate = -0.06, $SE = 0.02$, $t(528) = -3.86$, $p = .03$). Thus, only the experimental group improved on T1 level vs. T4 falling.

3. Discussion

This study investigated lexical tone contour categorization and discrimination by English listeners before and after minimal-tone-pair word training with only talker variability (control) or L2-Mandarin regional accent as well as talker variability (experimental). As predicted, accent variability facilitated post-training tone categorization and discrimination.

English listeners were required to assimilate Mandarin tones to tone contour icons due to the lack of lexical tones in their native phonological system. Before training, the experimental group showed Categorized assimilation of T1 level, which was Uncategorized by the control group. Both groups showed correctly Categorized T2 rising and T3 dipping, but Uncategorized or incorrectly Categorized assimilation of T4 falling. Their contrast assimilations were thus Two-Category for T1 level vs. T2 rising, T3 dipping, or T4 falling in experimental group, but T2 rising vs. T3 dipping for both groups. T1 level vs. T2 rising or T3 dipping in control group, T4 falling vs. T2 rising or T3 dipping in experimental group were Uncategorized-Categorized. T1 level vs. T4 falling was Single-Category for controls. Both groups discriminated all contrasts very well before training except for T1 level vs. T4 falling in control group, consistent with PAM predictions for these contrast assimilation types. Although lacking tones in their native phonological system, English listeners are *not* deaf to f_0 height and contour [5] as nonspeech patterns represented by icons. Moreover, minimal-tone-contrast word training with high accent and talker variability enhanced performance more than talker variability alone did.

After training, experimental condition Categorized all tones correctly, presumably as L2 phonological components. Correspondingly, their tone discriminations became equally excellent – their lower sensitivity to differences between T1 level and T4 falling than to other tone pairs prior to training disappeared after training. Listeners in the control group shifted their assimilation of T1 level from Uncategorized to Categorized. However, they failed to correctly Categorize T4 falling after training, although they did shift from incorrectly Categorizing it as T1 level to splitting their choices equally between falling and level icons, i.e., Uncategorized. Thus for them, T4 falling vs. T1 level became Uncategorized-Categorized after training. Their discrimination of this contrast remained poorer than for their Two-Category assimilation of T1 level vs. T3 dipping, again consistent with PAM principles.

Listeners in control condition improved their T1 level categorization to Categorized assimilation after minimal-tone-contrast word training, indicating that high talker variability alone can facilitate tone perception, which is in line with prior studies, such as[12], [15]. In addition to high talker variability, listeners in the experimental condition received high accent variability, which aided them forming T4 falling category. More importantly, their tone contour categorization and discrimination suggest that they had established L2 tone phonological categories for the tones, with excellent discrimination of all tone contrasts.

4. References

[1] Y. R. Chao, *Mandarin primer: An intensive course in spoken Chinese*. MA: Harvard University Press, 1948, p. 336.

[2] A. Abramson, "The tones of central Thai: Some perceptual experiments," in *Studies in Tai linguistics*, J. G. Harris and J. Chamberlain, Eds. Bangkok: Central Institute of English Language, 1975, pp. 1–16.

[3] C. K. So and C. T. Best, "Categorizing Mandarin tones into listeners' native prosodic categories: The role of phonetic properties," *Poznan Studies in Contemporary Linguistics*, vol. 47, no. 1, pp. 133–145, 2011, doi: 10.2478/psicl-2011-0011.

[4] C. K. So and C. T. Best, "Phonetic influences on English and French listeners' assimilation of Mandarin tones to native prosodic categories," *Stud Second Lang Acquis*, vol. 36, no. 2, pp. 195–221, Jun. 2014, doi: 10.1017/S0272263114000047.

[5] P. A. Hallé, Y.-C. Chang, and C. T. Best, "Identification and discrimination of Mandarin Chinese tones by Mandarin Chinese vs. French listeners," *Journal of Phonetics*, vol. 32, pp. 395–421, 2004, doi: 10.1016/S0095-4470(03)00016-0.

[6] C. Kiriloff, "On the auditory perception of tones in Mandarin," *Phonetica*, vol. 20, pp. 63–67, 1969, doi: 10.1159/000259274.

[7] H. Bluhme and R. Burr, "An audio-visual display of pitch for teaching Chinese tones," *Stud. Linguistics*, vol. 22, pp. 51–57, 1971.

[8] X. S. Shen, "Toward a register approach in teaching Mandarin tones," *Journal of the Chinese Language Teachers Association*, vol. 24, pp. 27–47, 1989.

[9] C. K. So and C. T. Best, "Cross-language perception of non-native tonal contrasts: Effects of native phonological and phonetic influences," *Language and Communication*, vol. 53, no. Pt 2, pp. 273–293, 2010, doi: 10.1111/j.1743-6109.2008.01122.x.Endothelial.

[10] Y. Wang, M. M. Spence, A. Jongman, and J. A. Sereno, "Training American listeners to perceive Mandarin tones," *The Journal of the Acoustical Society of America*, vol. 106, no. 6, pp. 3649–3658, 1999, doi: 10.1121/1.428217.

[11] K. Zhang, G. Peng, Y. Li, J. W. Minett, and W. S.-Y. Wang, "The effect of speech variability on tonal language speakers' second language lexical tone learning," *Frontiers in Psychology*, vol. 9, p. 1982, 2018, doi: 10.3389/fpsyg.2018.01982.

[12] S. Wiener, M. K. M. Chan, and K. Ito, "Do explicit instruction and high variability phonetic training improve nonnative speakers' Mandarin tone productions?," *The Modern Language Journal*, vol. 104, no. 1, pp. 152–168, 2020, doi: 10.1111/modl.12619.

[13] K. Johnson, "Resonance in an exemplar-based lexicon: The emergence of social identity and phonology," *Journal of Phonetics*, vol. 34, no. 4, pp. 485–499, 2006, doi: 10.1016/j.wocn.2005.08.004.

[14] J. Barcroft and M. S. Sommers, "Effects of acoustic variability on second language vocabulary learning," *Stud. Sec. Lang. Acq.*, vol. 27, no. 03, 2005, doi: 10.1017/S0272263105050175.

[15] P. C. M. Wong and T. K. Perrachione, "Learning pitch patterns in lexical identification by native English-speaking adults," *Applied Psycholinguistics*, vol. 28, no. 4, pp. 565–585, 2007, doi: 10.1017/S0142716407070312.

[16] T. Laméris and B. Post, "The combined effects of L1-specific and extralinguistic factors on individual performance in a tone categorization and word identification task by English-L1 and Mandarin-L1 speakers," *Second Language Research*, p. 026765832210900, Apr. 2022, doi: 10.1177/02676583221090068.

[17] Y. Li, C. T. Best, M. D. Tyler, and D. Burnham, "Regionally accented Mandarin lexical tones," *The Journal of the Acoustical Society of America*, vol. 148, no. 4, pp. 2474–2475, 2020, doi: 10.1121/1.5146856.

[18] Y. Li, C. T. Best, M. D. Tyler, and D. Burnham, "Tone variations in regionally accented Mandarin," in *Proceedings of the 21st Annual Conference of the International Speech Communication Association*, 2020, pp. 4158–4162. doi: 10.21437/interspeech.2020-1235.

[19] C. T. Best and M. D. Tyler, "Nonnative and second-language speech perception: Commonalities and complementarities," in *Second language speech learning: The role of language experience in speech perception and production*, M. J. Munro and O.-S. Bohn, Eds. Amsterdam: John Benjamins, 2007, pp. 13–34.

[20] C. T. Best, "A direct realist view of cross-language speech perception," in *Speech Perception and Linguistic Experience: Issues in Cross-Language Research*, W. Strange, Ed. Timonium, MD: York Press, 1995, pp. 171–204.

[21] J. Chen, C. T. Best, and M. Antoniou, "Native phonological and phonetic influences in perceptual assimilation of monosyllabic Thai lexical tones by Mandarin and Vietnamese listeners," *Journal of Phonetics*, vol. 83, p. 101013, 2020, doi: 10.1016/j.wocn.2020.101013.

[22] A. Reid *et al.*, "Perceptual assimilation of lexical tone: The roles of language experience and visual information," *Atten Percept Psychophys*, vol. 77, no. 2, pp. 571–591, 2015, doi: 10.3758/s13414-014-0791-3.

[23] A. R. Bowles, C. B. Chang, and V. P. Karuzis, "Pitch ability as an aptitude for tone learning: An aptitude for Tone," *Language Learning*, vol. 66, no. 4, pp. 774–808, Dec. 2016, doi: 10.1111/lang.12159.

[24] Q. Cai and M. Brysbaert, "SUBTLEX-CH: Chinese word and character frequencies based on film subtitles," *PLoS ONE*, vol. 5, no. 6, pp. e10729–e10729, 2010, doi: 10.1371/journal.pone.0010729.

[25] W. J. B. van Heuven, P. Mandera, E. Keuleers, and M. Brysbaert, "Subtlex-UK: A new and improved word frequency database for British English," *Quarterly Journal of Experimental Psychology*, vol. 67, no. 6, pp. 1176–1190, 2014, doi: 10.1080/17470218.2013.850521.

[26] R Core Team, *R: The R Project for Statistical Computing*. 2021. [Online]. Available: https://www.r-project.org/

[27] C. T. Best, G. W. McRoberts, and E. Goodell, "Discrimination of non-native consonant contrasts varying in perceptual assimilation to the listener's native phonological system," *The Journal of the Acoustical Society of America*, vol. 109, no. 2, pp. 775–794, 2001, doi: 10.1121/1.1332378.

[28] M. M. Faris, C. T. Best, and M. D. Tyler, "An examination of the different ways that non-native phones may be perceptually assimilated as uncategorized," *The Journal of the Acoustical Society of America*, vol. 139, no. 1, pp. EL1–EL5, 2016, doi: 10.1121/1.4939608.

[29] D. Bates, M. Mächler, B. Bolker, and S. Walker, "Fitting linear mixed-effects models using lme4," *ArXiv e-prints*, vol. arXiv:1406, 2014, doi: 10.18637/jss.v067.i01.

[30] U. Halekoh and S. Højsgaard, "A Kenward-Roger approximation and parametric bootstrap methods for tests in linear mixed models-the R package pbkrtest," *Journal of Statistical Software*, vol. 59, no. 9, pp. 1–32, 2014.

[31] J. Fox and S. Weisberg, *An R Companion to Applied Regression*, 3rd edition. SAGE Publications, Inc, 2018.

[32] R. V. Lenth, "Least-squares means: The R package lsmeans," *Journal of Statistical Software*, vol. 69, no. 1, Art. no. 1, 2016, doi: 10.18637/jss.v069.i01.

[33] J. Zhang and S. T. Mueller, "A note on ROC analysis and non-parametric estimate of sensitivity," *Psychometrika*, vol. 70, no. 1, pp. 203–212, 2005, doi: 10.1007/s11336-003-1119-8.

[34] N. A. Macmillan and C. D. Creelman, *Detection theory: A user's guide*, 2nd ed. Mahwah, N.J.: Lawrence Erlbaum Associates, 2005.

Accuracy-latency association in discrimination of L2 vowel contrasts

Yizhou Wang[a], Rikke L. Bundgaard-Nielsen[b], Brett J. Baker[a], Olga Maxwell[a]

[a]School of Languages and Linguistics, the University of Melbourne
[b]MARCS Institute for Brain, Behaviour and Development, University of Western Sydney
yizwang3@unimelb.edu.au, rikkelou@gmail.com, bjbaker@unimelb.edu.au
omaxwell@unimelb.edu.au

Abstract

This study investigates the efficacy of using response time-based metrics for assessing difficulty levels in L2 vowel discrimination. L1 Mandarin listeners discriminated six Australian English vowel contrasts with different cross-linguistic phonological statuses (cross-boundary vs within-category, according to PAM [1]). Results suggest that latency metrics, similarly to accuracy measures, can indicate the level of difficulty in AXB discrimination. Results also show that the correlation between accuracy and latency metrics is conditioned by the phonological status of the L2 vowel contrasts being tested. Strong accuracy-latency associations exist in cross-boundary contrasts, but no clear correlations are found for within-category pairs.

Index Terms: vowel discrimination, latency, accuracy

1. Introduction

It is well known that second language (L2) listeners at times encounter difficulty in discriminating target language phones that are non-contrastive in their native (L1) phonology inventory, e.g., it is notoriously difficult for Japanese listeners to discriminate the English /r/-/l/ contrast [2], [3]. This difficulty can be explained in terms of patterns of cross-language assimilation of L2 phones, according to the Perceptual Assimilation Model (PAM) [1] and its extension for L2 listening (PAM-L2) [4]. The central premise of PAM/PAM-L2 is that an L2 phone will be assimilated, or phonologically mapped, to an L1 category that is perceptually similar. For a pair of L2 phones, discrimination can be particularly difficult if both phones are perceived as equally good exemplars of a single L1 category, forming a Single-category (SC) pair. Discrimination can be relatively difficult when both L2 phones are perceived as exemplars of a single L1 category, but with different levels of phonetic similarity, i.e., they form a Category-goodness (CG) pair. If a pair of L2 phones are perceived as exemplars of two distinct L1 categories, they form a Two-category (TC) assimilation pair, for which good discrimination performance is predicted. Lastly, if one L2 phone is assimilated to an L1 category, but the other is not successfully mapped to any L1 category, they form an Uncategorised-categorised (UC) pair, which should also lead to good discrimination performance. TC and UC pairs thus represent *cross-boundary* contrasts, while CG and SC pairs represent *within-category* differences. In general, the discriminability of L2 contrasts follows a hierarchy based on PAM/PAM-L2 types, such that TC/UC > CG > SC [5].

While PAM/PAM-L2 has been shown to be a successful model in predicting the discriminability of non-native phones by analysing cross-language category mapping, support for the model has come primarily from studies relying on accuracy measures, i.e., percent correct (Pc), whilst some PAM-based studies also use latency measures as additional measures [6]–[8]. Accuracy is, however, an offline metric indicating end-point performance, and it cannot capture difficulty encountered during online processing, except as an inference.

To address this shortcoming, theories such as Automatic Selective Perception (ASP) [9], [10] call for the inclusion of latency measures such as response time (RT) and event-related potential (ERP) as additional indices for evaluating perceptual performance. In particular, the ASP model posits that automatic and attentional processing constitute a continuum of "effort", which can be defined in terms of response latency measures, e.g., mean response time (MRT). In particular, the ASP model predicts that discriminating within-category contrasts (i.e., CG/SC in PAM) requires a mode of processing which is relatively slow and susceptible to adverse listening conditions. In contrast, discriminating cross-boundary pairs (i.e., TC/UC in PAM) uses a fast mode of speech processing, which is robust even in adverse conditions.

In addition, computational cognition research [11]–[13] suggests response time variability (RTV, which is defined as the intra-participant standard deviation of response times) is also an important metric for evaluating discrimination performance. While MRT captures the central tendency of discrimination RT, RTV reflects perceptual instability during the decision-making process. In the present study, we aim to explore the association of three discrimination performance metrics, including accuracy (Pc), automaticity (MRT), and instability (RTV) in L1 Mandarin listeners' discrimination of six Australian English (AusE) vowel contrasts, /æ/-/ɐ/, /ɐː/-/ɐ/, /iː/-/ɪ/, /æ/-/e/, /eː/-/æ/, and /ɜː/-/e/, which potentially form different PAM/PAM-L2 contrasts. For instance, Mandarin has only one low vowel /a/, which can have various allophones including [ɐː] (in open syllables), [ɐ] (in closed syllables), [æ] (after /j/ and before /n/) [14]. Therefore, AusE contrasts /æ/-/ɐ/ and /ɐː/-/ɐ/ can be difficult for Mandarin listeners because these vowels are not contrastive phones in Mandarin. Similarly, /iː/-/ɪ/ can also be challenging since Mandarin does not have duration contrasts [14]. Additionally, Mandarin does not have mid front monophthongs /e/ or /eː/, and therefore AusE /e/ and /eː/ can be deemed as "unfamiliar" vowels for Mandarin listeners. Lastly, Mandarin has a mid-vowel /ɤ/ which is potentially similar to AusE /ɜː/ in perception. Therefore, the six contrasts cover a wide range of phonological scenarios involving both "familiar" and "unfamiliar" vowels for Mandarin listeners. More specifically, this paper aims to address two research questions:

(1) Can response latency measures (MRT and RTV), like accuracy (Pc), serve as effective indices of discrimination

difficulty for L2 vowel pairs of different PAM/PAM-L2 assimilation types?

(2) How are accuracy and latency metrics correlated with each other across and within different PAM/PAM-L2 assimilation pairs?

2. Methods

2.1. Participants

The participants were twenty native Mandarin speakers (M_{age} = 26.4, *SD* = 3.2; 17 females), who were post-graduate students at an Australian university with a mean Length of Residence (LoR) in Australia of 1.5 years (*SD* = 1.4). All spoke English as an L2 and had an average English learning history of 15.6 years (*SD* = 4.0) in classroom-based settings prior to their arrival in Australia. All participants originally came from PR China, and all were native speakers of Standard Mandarin [14]. None reported fluency in a third language. Research suggests that L2 speech proficiency is influenced by language dominance [15], i.e., the relative linguistic command in each language, and therefore all participants completed a Bilingual Language Profile (BLP) questionnaire [16]. On average, their Mandarin dominance score (*M* = 200, *SD* = 10) was significantly higher than English (*M* = 90, *SD* = 14), *t* = 23.4, *p* < .0001, indicating that our participants were not balanced bilingual speakers.

2.2. Stimuli

Two female native speakers of AusE were recorded producing the eight relevant target vowels in /hVba/ carrier pseudowords, which minimise coarticulatory effects and provide a controlled phonological structure [17]–[19]. The speakers were instructed to produce the vowels in a clear citation style [17], [20], [21], and they produced each stimulus pseudoword multiple times. All stimuli were inspected in the phonetic software Praat [22], and the "best" instances were selected to be used in the perception experiments on the basis of voice and phonetic quality. The average acoustic measurements are reported in Table 1.

Table 1. *Acoustic properties of the vowel stimuli.*

AusE V	Dur (ms)	F1 (Hz)	F2 (Hz)	F3 (Hz)
/ɐː/	221	841	1247	3168
/ɐ/	78	894	1313	3037
/æ/	110	793	1616	3133
/eː/	225	645	2208	3152
/e/	88	638	2105	2968
/ɜː/	207	519	1626	2827
/iː/	171	441	2779	3378
/ɪ/	72	465	2481	3276

2.3. Experimental design

To answer the research questions, we designed two tasks. The first experiment was a perceptual assimilation task, which examines L2-to-L1 similarities, where the results will inform the PAM/PAM-L2 type for each vowel contrast. In this task, participants first watched an instruction video explaining the purpose and structure of the task. They were then instructed to complete the task in a quiet environment wearing headphones. The experiment consisted of 80 randomised trials (eight vowels × two speakers × five repetitions). On each trial, participants first heard an AusE /hVba/ token and were then asked to categorise the target vowel into a native language (Mandarin) category. We adopted the *whole-system* approach [17], [18] and provided all phonotactically attested L1 categories as categorisation options, including all Mandarin monophthongs, diphthongs, and triphthongs: /a, i, y, u, o, ɤ, wə, yə, jə, ja, wa, aj, aw, əw, əj, jaw, waj, iəw, wəj/. Responses were made with the help of Mandarin keywords labelled on a virtual keyboard with *Pinyin* orthography. After each categorisation, participants heard the stimulus again and were prompted to provide a goodness rating based on the perceived similarity between the stimulus and the response on a seven-point scale, where 1 = "very different", 7 = "very similar".

The second experiment was an AXB discrimination task, which is the conventional paradigm for testing discrimination in studies based on PAM/PAM-L2 [5], [18], [23]. Six AusE vowel pairs, /æ/-/ɐ/, /ɐː/-/ɐ/, /iː/-/ɪ/, /æ/-/e/, /eː/-/æ/, and /ɜː/-/e/, were tested. On each trial, participants heard three consecutive /hVba/ stimuli and were asked to respond whether the middle stimulus (X) was more similar to the first stimulus (A) or the third one (B). In each triplet, A and B represented two different L2 vowel categories while X was phonologically identical to either A or B. Participants made responses by pressing the key "F" (X = A) or key "J" (X = B) on their keyboard. We presented participants with four counterbalanced triplet types (AAB, ABB, BAA, and BBA), ensuring that the participants were unable to predict the correct answer. In all trials, stimulus A and B were produced by one speaker, and X produced by the other. The task had 192 trials in total (6 pairs × 4 triplets × 2 speakers × 4 repetitions).

To ensure that the AXB task taps into phonological processing, rather than phonetic processing, we set the interstimulus interval (ISI) at 1000 ms [24]. On each trial, participants responded within 2000 ms, and response RTs were systematically measured from the offset of the last stimulus presentation to the time of response. Three metrics are calculated for each participant to assess their discrimination performance, including Pc (i.e., percentage correct), MRT (i.e., the average response time in correct trials), and RTV (i.e., the standard deviation of response times in correct trials). The experiments were developed and delivered using PsyToolkit 3.0 [25], [26].

3. Results

3.1. Assimilation results

The categorisation responses were averaged for participants and summarised in Table 2. By applying the 70% criterion in PAM-based studies [5], [17], [18], six out of eight AusE vowels were deemed as categorised: three AusE vowels, /ɐː/, /ɐ/, and /æ/, were consistently categorised as instances of the Mandarin low vowel /a/; AusE vowels /iː/ and /ɪ/ were consistently categorised as instances of the Mandarin high front vowel /i/, and AusE /ɜː/ was categorised as an instance of the Mandarin mid vowel /ɤ/. AusE /eː/ and /e/ were not assimilated into any Mandarin vowel category consistently, leaving them as uncategorised phones. Both were perceived as similar to Mandarin vowels /ej/ and /aj/, and to a lesser degree, /je/ and /ɤ/.

The assimilation status of individual vowels determines the PAM/PAM-L2 types of each vowel pair in /æ/-/ɐ/, /ɐː/-/ɐ/, /iː/-/ɪ/, /æ/-/e/, /eː/-/æ/, and /ɜː/-/e/. Since the vowels in the first three pairs were categorised into a single Mandarin vowel, they form either an SC or CG pair, depending on whether there is a perceptible difference in the goodness rating measure. Both

AusE /æ/ and /ɐ/ were perceived as relatively poor exemplars of Mandarin /a/ with a mean goodness score of 4.75 (*SD* = 1.21) and 4.78 (*SD* = 1.17), respectively. We built a linear mixed-effects model (LMM) at the decision level with random slopes and intercepts set for all participants. Then, a Wald Chi-squared test compared the mean scores, and there was no significant difference, $\chi^2(1) = 0.092$, $p = .762$. Therefore, we deemed /æ/-/ɐ/ as an SC pair.

Table 2. *Assimilation matrix of the AusE vowels: Percentage of responses.*

V	/a/	/i/	/ɤ/	/ej/	/aj/	/je/
/ɐ:/	96	1	3			
/ɐ/	96		2		1	
/æ/	76	1	3	5	14	
/e:/	1	1	10	46	30	12
/e/	4	1	9	45	33	8
/ɜ:/			99		1	
/i:/		87	2	9		1
/ɪ/		73	3	19		5

Next, AusE /ɐ:/ was perceived as a relatively good exemplar of Mandarin /a/ with a mean goodness score of 5.49 (*SD* = 1.04). For /ɐ:/-/ɐ/, we built another LMM which revealed a significant difference of goodness, $\chi^2(1) = 10.873$, $p < .001$. We therefore deemed /ɐ:/-/ɐ/ as a CG pair. Lastly, while both AusE /i:/ and /ɪ/ were both categorised as Mandarin /i/, the former was perceived as a better exemplar than the latter (mean goodness = 5.04, *SD* = 1.16, and mean goodness = 4.10, *SD* = 1.25, respectively). When analysed with an LMM, the difference was significant, $\chi^2(1) = 13.438$, $p < .001$. Therefore, /i:/-/ɪ/ formed another CG pair. For the last three pairs (/æ/-/e/, /e:/-/æ/, and /ɜ:/-/e/), the assimilation patterns resulted in one uncategorised vowel (/e:/ or /e/) and one categorised vowel (/æ/ or /ɜ:/), and thus they formed three UC pairs. Based on the prediction that discrimination accuracy should form a hierarchy that TC/UC > CG > SC [5], we predict that Mandarin listeners will show good discrimination performance in AusE /æ/-/e/$_{UC}$, /e:/-/æ/$_{UC}$, and /ɜ:/-/e/$_{UC}$, relatively good discrimination performance in AusE /ɐ:/-/ɐ/$_{CG}$ and /i:/-/ɪ/$_{CG}$, and poor performance in AusE /æ/-/ɐ/$_{SC}$.

3.2. Discrimination results

The listeners' performance in the AXB discrimination task is summarised in Table 3. The lowest accuracy was found in the SC pair /æ/-/ɐ/ (Pc = 61%), and the highest accuracy was found in the UC pair /ɜ:/-/e/ (Pc = 88%). For all six pairs, the accuracy measure (Pc) differed significantly when checked by an LMM, $\chi^2(5) = 78.208$, $p < .001$. Similarly, we built an LMM for MRT, which also revealed a significant effect of contrast, $\chi^2(5) = 81.809$, $p < .001$. We similarly found a significant effect for RTV, $\chi^2(5) = 45.55$, $p < .001$. To summarise, all AusE vowel pairs differed significantly in terms of the three metrics obtained.

A series of Tukey-adjusted *post hoc* tests (see Table 4) revealed significant differences between AusE vowel pairs of different PAM/PAM-L2 contrast types. Additionally, the three metrics, i.e., Pc, MRT, and RTV, showed some level of both commonality and complementarity: All three metrics detected a significant difference between /æ/-/ɐ/$_{SC}$ and /e:/-/æ/$_{UC}$, /æ/-/ɐ/$_{SC}$ and /ɜ:/-/e/$_{UC}$, /ɐ:/-/ɐ/$_{CG}$ and /ɜ:/-/e/$_{UC}$, and between /i:/-/ɪ/$_{CG}$ and /ɜ:/-/e/$_{UC}$. In addition, the Pc metric captured differences that were not shown in the MRT and RTV patterns, between /æ/-/ɐ/$_{SC}$ and /i:/-/ɪ/$_{CG}$, /æ/-/ɐ/$_{SC}$ and /æ/-/e/$_{CG}$, and between /ɐ:/-/ɐ/$_{CG}$ and /æ/-/e/$_{UC}$. The MRT metric uniquely captured significant differences between /ɐ:/-/ɐ/$_{CG}$ and /i:/-/ɪ/$_{CG}$, /i:/-/ɪ/$_{CG}$ and /e:/-/æ/$_{UC}$, and between /i:/-/ɪ/$_{CG}$ and /æ/-/e/$_{UC}$, which the Pc metric did not capture. Similarly, RTV captured the difference between /ɐ:/-/ɐ/$_{CG}$ and /e:/-/æ/$_{UC}$, /i:/-/ɪ/$_{CG}$ and /e:/-/æ/$_{UC}$, and between /i:/-/ɪ/$_{CG}$ and /æ/-/e/$_{UC}$, where the Pc metric did not show a significance. Importantly, all observed significant differences were consistent with the prediction of discrimination performance based on PAM/PAM-L2, such that TC/UC > CG > SC [5].

Table 3. *AXB discrimination results. Standard deviations are shown in parentheses.*

Pair	PAM	Pc (%)	MRT	RTV
/æ/-/ɐ/	SC	61 (15)	633 (203)	382 (77)
/ɐ:/-/ɐ/	CG	71 (13)	565 (167)	387 (118)
/i:/-/ɪ/	CG	74 (11)	693 (183)	417 (98)
/æ/-/e/	UC	79 (11)	467 (127)	289 (83)
/e:/-/æ/	UC	83 (11)	500 (176)	327 (100)
/ɜ:/-/e/	UC	88 (13)	528 (158)	290 (118)

Table 4. *P-values in pairwise comparisons. Significant comparisons are set in bold face.*

Contrast	1	2	3	4	5
1 /æ/-/ɐ/$_{SC}$	-				**Pc**
2 /ɐ:/-/ɐ/$_{CG}$	.065	-			
3 /i:/-/ɪ/$_{CG}$	**.007**	.973	-		
4 /æ/-/e/$_{UC}$	**< .001**	.138	.519	-	
5 /e:/-/æ/$_{UC}$	**< .001**	**.007**	.065	.887	-
6 /ɜ:/-/e/$_{UC}$	**< .001**	**< .001**	**.001**	.121	.668
1 /æ/-/ɐ/$_{SC}$	-				**MRT**
2 /ɐ:/-/ɐ/$_{CG}$	.394	-			
3 /i:/-/ɪ/$_{CG}$	.553	.007	-		
4 /æ/-/e/$_{UC}$	**< .001**	.074	**< .001**	-	
5 /e:/-/æ/$_{UC}$	**.004**	.460	**< .001**	.935	-
6 /ɜ:/-/e/$_{UC}$	**< .001**	**.003**	**< .001**	.888	.338
1 /æ/-/ɐ/$_{SC}$	-				**RTV**
2 /ɐ:/-/ɐ/$_{CG}$	.999	-			
3 /i:/-/ɪ/$_{CG}$	.752	.856	-		
4 /æ/-/e/$_{UC}$	**.005**	**.003**	**< .001**	-	
5 /e:/-/æ/$_{UC}$	.262	.177	**.008**	.670	-
6 /ɜ:/-/e/$_{UC}$	**.006**	**.003**	**< .001**	.999	.688

3.3. Correlation between performance metrics

The second aim of the present study is to analyse how and to what extent the three implemented discrimination metrics (Pc, MRT, and RTV) correlate with each other. To address this, we carried out a series of Pearson's *r* correlation analyses based on the standardised score (*z*-score) of the three metrics within each vowel pair and across different vowel pairs, see Table 5. The results showed that the two latency-based metrics, MRT and RTV, were positively and significantly (p's < .01) correlated in four out of six vowel pairs (except /æ/-/ɐ/$_{SC}$ and /e:/-/æ/$_{UC}$). When the vowel pairs were pooled together based on phonological status, both within-category pairs (SC/CG) and cross-boundary pairs (UC) showed a very high correlation coefficient ($r = .595$ and $r = .648$, respectively, p's < .001). When all six AusE vowel pairs were pooled together, the correlation between MRT and RTV was still very high ($r = .695$, $p < .001$). In other words, the MRT-RTV correlation

tended to be similarly robust for within-category (SC/CG) as well as cross-boundary pairs (UC).

Table 5. *Correlation between metrics.*

Contrast	Pc ~ MRT	Pc ~ RTV	MRT ~ RTV
1. /æ/-/ɐ/$_{SC}$	0.036	–.063	0.43
2. /ɐ:/-/ɐ/$_{CG}$	–.180	–.188	.696***
3. /i:/-/ɪ/$_{CG}$	0.166	0.423	.649**
4. /æ/-/e/$_{UC}$	–.539*	–.694***	.803***
5. /e:/-/æ/$_{UC}$	–.488*	–.098	0.243
6. /ɜ:/-/e/$_{UC}$	–.493*	–.459*	.697***
SC/CG pairs	0.003	0.057	.595***
UC pairs	–.520***	–.416***	.648***
All pairs	–.428***	–.347***	.695***

Surprisingly, our analyses revealed no significant correlation between Pc and MRT for the within-category pairs (p's > .05,) or when the SC and CG pairs combined (p > .05). However, the Pc-MRT correlation was significant for all three UC pairs, i.e., /æ/-/e/, /e:/-/æ/, and /ɜ:/-/e/ (r = –.539, –.488, and –.493, p = .014, .029, and .027, respectively). When the three UC pairs were pooled together, the Pc-MRT correlation was significant (r = –.520, p < .001). This suggests that the Pc-MRT correlation is significant for cross-boundary (UC) pairs but not within-category (SC/CG) pairs. Lastly, we found no significant correlations between Pc and RTV in the within-category pairs (p's > .05) nor for the SC and CG pairs combined (p > .05). But we did find significant Pc-RTV correlations for the UC pairs /æ/-/e/ (r = –.694, p < .001) and /ɜ:/-/e/ (r = –.459, p = .042). When all three UC pairs were pooled together, we again found a significant Pc-RTV correlation (r = –.416, p = .001). These results show that both Pc-MRT and Pc-RTV tend to be negatively correlated for cross-boundary (UC) pairs, but no clear association is found for within-category (SC/CG) pairs. In other words, the strength of accuracy-latency association in L2 vowel discrimination depends on the phonological status, i.e., the PAM/PAM-L2 type of the vowel pair in question.

4. General discussion

The present paper addresses a number of theoretical and methodological questions. Firstly, we examined whether discrimination latency measures such as MRT and RTV, like Pc, can serve as effective indices of discrimination difficulty for L2 vowel pairs of different PAM/PAM-L2 assimilation types. Our results suggest that all three metrics can be used to indicate the discrimination difficulty of L2 vowel discrimination, as predicted by the PAM/PAM-L2 theory [1], [4]. By analysing six AusE vowel contrasts and their pairwise comparisons (Table 4), the results suggest that the accuracy and latency metrics have commonalities and complementarities. Accuracy (Pc) can at times capture between-pair differences that the latency metrics (MRT and RTV) do not show a difference for, and conversely sometimes latency metrics detect differences that are not visible in the accuracy measure.

Generally, our results support the ASP model's [9], [10] prediction that L2 vowels of different phonological statuses, i.e., perceptual assimilation types, will lead to different latency patterns: L2 listeners show accurate and fast speech processing for cross-boundary contrasts, while discrimination of within-category pairs is more difficult and less automatic. Typically, SC and CG pairs tend to rely on L2 listeners directing their cognitive resources to the reminiscent phonetic details, which is deliberately discouraged by the long ISI value in the AXB task [24]. On the contrary, UC pairs, which represent cross-boundary contrasts, are sufficiently discriminated based on coarse-grained phonological coding, which is resistant to the decay of short-term sensory memory in a long ISI design [24].

We also explored the effectiveness of including RTV as an additional latency metric, as it is suggested in computational cognition research [11], [12] to be informative of neural instability during the discrimination process. Thus, our findings also extend the ASP model's premise that the cross-boundary L2 contrasts are processed in a more *stable* manner than within-category contrasts. Methodologically, the results suggest that discrimination research should use both accuracy and latency since they can tap into different aspects of discrimination performance. Clearly, the resolution of the accuracy measure depends on the number of trials. For instance, with 32 trials testing an L2 vowel pair in discrimination, we could obtain a resolution of 100%/32 or 3.125%, and when fewer trials were tested, we would obtain a more coarse-grained measure at the participant level. On the contrary, the latency metrics MRT and RTV are gradient in nature, and their baseline levels can be determined by a control condition that should not be difficult to listeners, e.g., a TC/UC pair.

The second research question asks how accuracy and latency metrics are associated and correlated with each other across and within different L2 vowel pairs. Our results suggest that the two latency-based metrics, MRT and RTV, tend to be positively correlated irrespective of the phonological status of the L2 vowel pair. This finding is consistent with previous research, which reports that MRT and RTV tend to covary in an approximately linear manner in two-choice tasks [27]. More strikingly, we observed that Pc-MRT and Pc-RTV correlations were conditioned by the phonological status of the L2 vowel pairs, i.e., their perceptual assimilation patterns, see Table 5. For within-category (SC/CG) pairs, we did not find significant correlations between accuracy and latency, but we found robust negative correlations for cross-boundary (UC) pairs, such that high accuracy measures accompany high response speed and low instability. This indicates that MRT and RTV in easy conditions, i.e., for cross-boundary (UC, potentially also TC) pairs, sufficiently reflect the cognitive demands of the decision process during an AXB discrimination task. But in difficult conditions, i.e., for within-category (CG and SC) pairs, they cannot effectively measure the cognitive demand.

Recall that the AXB task has counterbalanced speaker order and triplet types, and in case a listener responds randomly they will still achieve 50% accuracy. However, the latency measure estimated from these trials will reflect properties of guesses rather than the targeted cognitive process in discrimination procedures. Although it is difficult to tease apart guesses from the dataset based on behavioural measures, it is reasonable to assume that lower accuracy measures should accompany higher proportions of guesses. On the contrary, the proportion of guesses should be relatively low in easy conditions, i.e., for cross-boundary (UC) pairs, and MRT and RTV should effectively reflect the speed and stability of the discrimination process. This finding reveals a disadvantage of latency as compared to accuracy: Guesses due to an inability to discriminate a contrast will contaminate the latency metrics [28] evaluated from an AXB task. In general, accuracy and latency metrics can still complement each other, especially when a range of difficulty levels are tested: Accuracy might fail to capture more nuanced differences when participants tend to achieve ceiling level accuracy; On the contrary, latency might not effectively reflect cognitive demands when the accuracy is relatively low.

5. Acknowledgements

We want to thank the twenty participants for their time. Special thanks also go to Debbie Loakes and Catherine Roberts for their help in generating the stimuli.

6. References

[1] C. T. Best, "A direct realist view of cross-language speech perception," in *Speech perception and linguistic experience: Issues in cross-language research*, W. Strange, Ed. Timonium, MD: York Press, 1995, pp. 171–204.

[2] W. Strange and S. Dittmann, "Effects of discrimination training on the perception of/r-l/ by Japanese adults learning English," *Percept. Psychophys.*, vol. 36, no. 2, pp. 131–145, 1984.

[3] K. S. MacKain, C. T. Best, and W. Strange, "Categorical perception of English /r/ and /l/ by Japanese bilinguals," *Appl. Psycholinguist.*, vol. 2, pp. 369–390, 1981.

[4] C. T. Best and M. D. Tyler, "Nonnative and second-language speech perception: Commonalities and complementarities," in *Language experience in second language speech perception*, O.-S. Bohn, Ed. Amsterdam: John Benjamins, 2007, pp. 13–34.

[5] M. D. Tyler, C. T. Best, A. Faber, and A. G. Levitt, "Perceptual assimilation and discrimination of non-native vowel contrasts," *Phonetica*, vol. 71, no. 1, pp. 4–21, 2014.

[6] P. A. Hallé and C. T. Best, "Dental-to-velar perceptual assimilation: A cross-linguistic study of the perception of dental stop+/l/ clusters," *J. Acoust. Soc. Am.*, vol. 121, no. 5, pp. 2899–2914, 2007.

[7] C. T. Best and P. A. Hallé, "Perception of initial obstruent voicing is influenced by gestural organization," *J. Phon.*, vol. 38, no. 1, pp. 109–126, 2010.

[8] A. Kilpatrick, R. L. Bundgaard-Nielsen, and B. J. Baker, "Japanese co-occurrence restrictions influence second language perception," *Appl. Psycholinguist.*, vol. 40, no. 2, pp. 585–611, 2019.

[9] W. Strange and V. L. Shafer, "Speech perception in second language learners," in *Phonology and second language acquisition*, J. G. Hansen-Edwards and M. L. Zampini, Eds. Amsterdam: Benjamins, 2008, pp. 153–192.

[10] W. Strange, "Automatic selective perception (ASP) of first and second language speech: A working model," *J. Phon.*, vol. 39, no. 4, pp. 456–466, 2011.

[11] R. Ratcliff, "A theory of memory retrieval," *Psychol. Rev.*, vol. 85, pp. 59–108, 1978.

[12] R. Ratcliff, P. L. Smith, S. D. Brown, and G. McKoon, "Diffusion Decision Model: Current issues and history," *Trends Cogn. Sci.*, vol. 20, no. 4, pp. 260–281, 2016.

[13] R. Ratcliff and G. McKoon, "The Diffusion Decision Model: Theory and data for two-choice decision tasks," *Neural Comput.*, vol. 20, no. 4, pp. 873–922, 2008.

[14] S. Duanmu, *The phonology of Standard Chinese*. Oxford, UK: Oxford University Press, 2007.

[15] J. E. Flege, I. R. A. MacKay, and T. Piske, "Assessing bilingual dominance," *Appl. Psycholinguist.*, vol. 23, no. 4, pp. 567–598, 2002.

[16] D. Birdsong, L. M. Gertken, and M. Amengual, "Bilingual Language Profile: An easy-to-use instrument to assess bilingualism. COERLL, University of Texas at Austin, TX," *https://sites.la.utexas.edu/bilingual/*, 2012.

[17] R. L. Bundgaard-Nielsen, C. T. Best, and M. D. Tyler, "Vocabulary size matters: The assimilation of second language Australian English vowels to first-language Japanese vowel categories," *Appl. Psycholinguist.*, vol. 32, no. 1, pp. 51–67, 2011.

[18] R. L. Bundgaard-Nielsen, C. T. Best, and M. D. Tyler, "Vocabulary size is associated with second-language vowel perception performance in adult learners," *Stud. Second Lang. Acquis.*, vol. 22, pp. 433–461, 2011.

[19] W. Strange, A. Weber, E. S. Levy, V. Shafiro, M. Hisagi, and K. Nishi, "Acoustic variability within and across German, French, and American English vowels: Phonetic context effects," *J. Acoust. Soc. Am.*, vol. 122, no. 2, pp. 1111–1129, 2007.

[20] A. R. Bradlow and T. Bent, "The clear speech effect for non-native listeners," *J. Acoust. Soc. Am.*, vol. 112, no. 1, pp. 272–284, 2002.

[21] M. A. Picheny, N. I. Durlach, and L. D. Braida, "Speaking clearly for hard of hearing II: Acoustic characteristics of clear and conversational speech," *J. Speech, Lang. Hear. Res.*, vol. 29, no. 4, pp. 434–446, 1986.

[22] P. Boersma, "Praat, a system for doing phonetics by computer," *Glot Int.*, vol. 5, no. 9–10, pp. 341–345, 2001.

[23] C. T. Best, G. W. McRoberts, and E. Goodell, "Discrimination of non-native consonant contrasts varying in perceptual assimilation to the listener's native phonological system," *J. Acoust. Soc. Am.*, vol. 109, no. 2, pp. 775–794, 2001.

[24] J. F. Werker and J. S. Logan, "Cross-language evidence for three factors in speech perception," *Percept. Psychophys.*, vol. 37, pp. 35–44, 1985.

[25] G. Stoet, "PsyToolkit: A software package for programming psychological experiments using Linux," *Behav. Res. Methods*, vol. 42, no. 4, pp. 1096–1104, 2010.

[26] G. Stoet, "PsyToolkit: A novel web-based method for running online questionnaires and reaction-time experiments," *Teach. Psychol.*, vol. 44, no. 1, pp. 24–31, 2017.

[27] E. J. Wagenmakers, R. P. P. P. Grasman, and P. C. M. Molenaar, "On the relation between the mean and the variance of a diffusion model response time distribution," *J. Math. Psychol.*, vol. 49, no. 3, pp. 195–204, 2005.

[28] R. Ratcliff and F. Tuerlinckx, "Estimating parameters of the diffusion model: Approaches to dealing with contaminant reaction times and parameter variability," *Psychon. Bull. Rev.*, vol. 9, no. 3, pp. 438–481, 2002.

Something borrowed, something new: acquiring unexploited sets of feature contrasts

Rikke L. Bundgaard-Nielsen[1,2], Brett J. Baker[3] & Carmel O'Shannessy[2]

[1]MARCS Institute for Brain, Behaviour and Development; Western Sydney University; [2]School of Literature, Languages and Linguistics, Australian National University; [3]School of Languages and Linguistics, University of Melbourne

rikkelou@gmail.com; bjbaker@unimelb.edu.au; carmel.oshannessy@anu.edu.au

Abstract

The present paper addresses the question of whether Second Language (L2) segmental acquisition can be successful if the learner's Native Language (L1) does not implement an entire 'tier' of phonetic/articulatory information exploited in the L2. The results from a re-framed analysis of perceptual discrimination studies with speakers of three Indigenous Australian languages (Wubuy; Kriol; Light Warlpiri) indicate that this appears possible only when L2 learners can leverage phonetic aspects of their consonantal inventories to acquire some new featural contrasts, and that distributional information can act as a key to at least some new contrastive features. The results also indicate that voicing, especially in fricatives, appears particularly challenging.

Index Terms: L2 perception, voicing, frication, feature theory, Perceptual Assimilation Model.

1. Learning something truly new

The present paper addresses the question of whether Second Language (L2) segmental acquisition can be successful if a learner's Native Language (L1) does not implement an entire 'tier' of phonetic/articulatory information exploited in the L2. The acquisition of novel L2 *phonemes*, and the acquisition of new *phonetic realisations* of shared L1-L2 phonemes has received ample attention in the psycholinguistic and applied linguistics literature. Typically, however, the target L2 phonemes and/or novel L2 phonetic realisations of interest exploit linguistic variables already in use in the learners' L1. As a consequence, many studies have focused on understanding the degree to which perceptual, acoustic, and articulatory overlaps and differences between L1 and L2 phoneme inventories and their phonetic realisations hinder or help L2 segmental acquisition; e.g. [1][2][3]. Such studies contribute to our understanding of those aspects of L2 acquisition that pertain to the *re-tuning* of the perceptual systems to already systematically controlled and manipulated linguistic variables. It does not, however, directly address questions of learnability of truly new variables for contrast maintenance, a question that is important in terms of applied linguistics, but also in terms of the fundamental question of what is learnable. In other words, if a novel linguistic tier is encountered for the first time in an L2, can it be acquired? And to what extent will learners be able to attune to speech information that they have not previously had use for? This question indeed goes to the heart of to what extent phonological features are really 'universal' in some sense, as proposed in [4].

The present paper addresses these questions by reviewing and reconsidering the results from a series of speech perception studies with what we argue are genuinely *novel* phonemic distinctions, relying on *novel* tiers of linguistic information (voicing; frication) with speakers of the traditional Australian language Wubuy, and the Australian contact varieties (Roper) Kriol and Light Warlpiri, in long-term and continuing language contact situations in the Northern Territory of Australia. The combined results indicate that speakers of these languages can leverage phonetic aspects of their consonantal inventories to acquire some new featural contrasts, but not all of them.

2. The Perceptual Assimilation Model, Articulatory Organs, and the acquisition of novel phonological contrasts

One of the leading theories of L1 segmental acquisition, the Perceptual Assimilation Model (PAM; [1][2]), and its extension to L2 acquisition PAM-L2 [3] has its basis in Direct Realism [5], positing that the objects of speech perception are articulatory gestures, and that acoustics are important only because they carry information about the articulatory gestures of speech. This information is complemented by information about the same gestures obtained through other modalities (eg., vision, touch). PAM posits that language learning is possible throughout the lifespan but shaped by linguistic experience.

PAM further relies on the principles of Articulatory Phonology [6][7][8][9] which posits that the lips, tongue tip, tongue body, tongue root, velum, and larynx, and the glottis constitute distinct and potentially independent articulatory organs, as each of these can be used to create an articulatory constriction without inducing a constriction in another. The related Articulatory Organ Hypothesis (AOH; [10][11]) predicts that consonants produced by different articulators should be discriminated better than consonants produced by the same articulator, providing a pre-experience tendency in segmental speech perception. It is not clear to which degree the AOH predicts that (innate?) knowledge of articulatory organs provide only an initial set of perceptual biases scaffolding language learning or whether these biases are lifelong. PAM has focused less on what might constitute innate biases in speech perception and more on the role of linguistic experience in the first year of life, as infants develop from universal to language specific listeners. According to PAM, linguistic experience allows infants begin to recognise that particular constellations of articulatory gestures contrast with other constellations, and from these constellation, they tune into the higher-order invariants that characterise their linguistic environment [12].

Taken together, however, the AOH and PAM raise questions of the learnability of L2 phonemes, and L2 phonological contrasts, that differ from the L1 in terms of the organs and articulatory gestures used—and importantly, in the phonetic *tiers*, *features* or (combinations of) *organs* that are used for contrast maintenance. Indeed, it may be possible to argue that: **(1)** spoken language is (biologically) special; **(2)** the articulatory organs have special status as an initial bootstrapping mechanism, at least in very early infancy; **(3)** linguistic experience allows for identification of language-specific use of the articulatory organs (and the potential 'neglect' of some organs as important-in-speech; and, **(4)** language acquisition may be a 'use it or lose it'-scenario with respect to the continued importance of each articulatory organ.

In other words, it is possible to hypothesise that organs—and importantly the (phonetic) tiers of information that they produce for the purposes of contrast maintenance—must have a role to play in the L1, if they are to maintain their initial, pre-experience, linguistic relevance. If a tier of information is not in use in the L1, L2 acquisition would presumably be difficult, as learners would have to identify information which is non-speech or irrelevant to speech in the L1 as important to contrast maintenance in the L2. This is different, in essence, to other challenges in L2 acquisition pertaining to resetting of, for instance, VOT boundaries, or carving a given corner of the vowel space into a different set of L2 vowels. These latter examples may constitute 'retuning' of existing speech perception/production skills, while the acquisition of new tiers of linguistic information demands the expansion of the perceptual space in relation to what kinds of major articulatory mechanisms can constitute speech at all.

Few studies have examined cross-language or L2 perception—by adults or infants—of non-native contrasts that can be argued to be truly new in the sense that they rely on the perception of new 'tier' of information. Studies of non-native perception of click consonants may provide one of the few exceptions: click consonant articulations are among the typologically rare ingressive phonemes. Clicks, however, do not present themselves as particularly difficult to discriminate for non-native listeners: Zulu clicks are discriminated with relative ease by adult speakers of English, and amplitude-modified Zulu clicks are discriminated as accurately as native English /ba/ versus /da/ by infants [13]. Clicks are however often perceived as non-speech mouth sounds, and it is possible that crosslinguistic discrimination success does not reflect linguistic processing (alone).

3. Discrimination of L2 stop voicing and fricatives by speakers of three Australian Indigenous languages

Traditional (i.e., non-contact) Australian Indigenous languages are quite homogenous in their phonological inventories [15]. They typically have 4-6 Places of Articulation (POAs) in stops, the same number in nasals, and several laterals, with the maximal inventories including bilabial, lamino-dental, apico-alveolar, apico-retroflex, lamino-palatal and dorso-velar POAs (see Table 1). Many Australian Indigenous languages also have multiple rhotics, but they rarely have voicing distinctions in stops (or anywhere else), though some have a contrast in stops for duration (sometimes described as fortis-lenis) with a duration ratio of approximately 1:2.5 [14]. Most Indigenous Australian languages have a complete absence of fricatives [15]. These inventories thus differ from that of Australian English in a number of critical ways: English employs a voicing-based distinction not just in stop consonants, but also in affricates and fricatives. To conserve space, we will burden the reader with the assumption that they are well-enough familiar with the phoneme inventory of English and abstain from including a table.

Table 1. *Maximal consonant inventory for Indigenous Australian languages.*

	Lab.	Dent.	Alv.	Retrofllex	Palatal	Velar
Stop	p	t̪	t	ʈ	c	k
Nasal	m	n̪	n	ɳ	ɲ	ŋ
Lateral		l̪	l	ɭ	ʎ	
Trill/ tap			ɾ			
Glide	w			ɻ	j	

Acoustically, the stop consonants of Indigenous Australian languages are often voiceless, including intervocalically, and (linguists') perceptions of the voicing characteristics of stops vary by language and context (and by linguist). These perceptions are reflected in the orthographies of Australian languages (often produced by Europeans).

Intervocalic stops are often characterised by extraordinarily long Voice Termination Times (VTT), the time from offset of the previous (nasal or vowel) segment until cessation of vocal fold vibration during the constriction phase of the consonant [16], particularly the lenis series. This suggests that speakers are in no rush to turn off vocal fold vibration as they transition into the stop, and, in fact, there is no real reason to do so, given that there is no voicing distinction to maintain in stop consonants. Long VTTs further suggest that the timing of vocal fold vibration is not intentionally controlled (stopped or extended). In languages with a long/short or fortis/lenis contrast, the long VTTs can give the impression of short/lenis stops being voiced while the long/fortis stops can be perceived as voiceless as physiological/articulatory constraints often result in voicing 'running out' in these long stops, particularly in velars. [16] gives an average duration of 22 ms for VTT in the fortis series in Bininj Gun-wok. Extended, continuous contact between speakers of Indigenous Australian languages (traditional as well as contact varieties) with English/English-based varieties provides a rare opportunity to study acquisition of entirely novel phonetic variables: stop voicing distinctions and a new Manner of Articulation (MOA) in fricatives that contrast not just in POA but also in voicing.

We present here summaries, re-framing, and re-interpretation of three studies of stop and fricative perception with speakers of Indigenous Australian languages: the English-lexified creole Roper Kriol; the mixed language Light Warlpiri, and the traditional Indigenous Australian language Wubuy. The studies have been published in [17] and [18] but the combined results and comparisons provide an opportunity to assess the outcomes on a more general level. Theoretically, this is important because the set of studies provides a rare departure from those which focus on the retuning of boundary placements and target values on shared already-exploited variables, as discussed in the introduction. The studies discussed here also depart from those which focus on what we might conceive of as 're-segmentation' of a particular variable, increasing or decreasing the number of boundaries on a shared and continuous variable in a continuous space, e.g, moving from two-to-three or three-to-two VOT boundaries, or increasing or decreasing the number of vowels in any given part of the vowel space. Similarly, we are not concerned with defining an 'end

state' of L2 acquisition: the contact environment is lifelong and changing over time and speaker generations.

The three studies reported on jointly here made use of a cross-speaker XAB discrimination paradigm (see [17], [18] for details). The stimulus materials were produced by three female Australian English speakers, and consisted of a number of stop contrasts, fricative contrasts and a stop-fricative contrast in a combination of syllable-initial and intervocalic position (See Table 2). In addition to testing discrimination of English stop contrasts /p b/ and /k g/ in intervocalic position, we also tested discrimination of Kriol-like stop contrasts in which the voiceless stop is characterised not only by a longer VOT, but also by a longer constriction duration (CD), consistent with the constriction duration found in the fortis/long stops in those of the Kriol substrate languages that have a fortis/lenis contrast. Discrimination of each target contrast was tested in blocked discriminations tasks consisting of 72 individual trials, with an ISI of 500 ms and an ITI 3000 ms. Missed trials were repeated.

The participants in the three studies were 11 speakers of Wubuy (age range 25-65), all from Numbulwar community in East Arnhem Land; 13 Light Warlpiri speakers (age range 16-33), all from Lajamanu Community; and 11 speakers of Kriol (age range (18-50), all from Numbulwar community also. All speakers were lifelong L2 learners/users of Australian English, the language of instruction in schools and some workplaces in their communities, as well as the language used in communication with members of the broader Australian community, including non-Indigenous people in their local communities. All participants had some English literacy, and some also had some literacy in their native languages. The Wubuy and Light Warlpiri speakers typically had some competence in Kriol, and most had at least some competence in one or more other Indigenous languages (particularly Warlpiri, in the case of the Light Warlpiri speakers).

4. Is acquisition happening?

Table 2 summarises the results from the studies introduced above.

Table 2. *Contrasts and results from Wubuy; Kriol; Light Warlpiri (LW) participants. /CV/ = initial; /CVC/ = intervocalic. * indicates no data from Wubuy and Kriol. % values = accuracy where performance was significantly above chance, but below ceiling.*

	/CV/	Result	/VCV/	Result
Practice	/p k/	**ALL**		
Stop-Stop	/p b/	LW ~70%	/p b/	LW ~65%
			/p: b/	LW + Kriol
	/k g/	**NONE**	/k g/*	LW ~65%
			/k: g/*	LW
Stop-Fricative	/b v/	**ALL**		
Fricative-Fricative	/s z/	**NONE**		
	/s ʃ/	**ALL**		

These results allow us to examine the acquisition success of truly new information tiers used for contrast maintenance in an L2 in two different domains: **(1)** acquisition of a new MOA in the case of frication, and **(2)** acquisition of voicing-based distinctions (in stops). It further allows us to discuss the difficulties that may arise from having to acquire not just one of the two new tiers (frication *and* voicing) for successful L2 acquisition, but in the case of fricative voicing contrasts, such as English /s z/, two novel tiers simultaneously.

4.1. Acquisition of a new MOA (frication)

The results relevant to the question of acquisition of a new MOA—frication—are consistent across the three learner groups. Speakers of the traditional language Wubuy as well as the two contact varieties Light Warlpiri and Kriol all successfully discriminate the POA-based fricative-fricative contrast /s ʃ/, and the MOA fricative-stop contrast /b v/, both in word-initial position. In the case of Kriol and Light Warlpiri speakers, performance likely reflect L1-like perception of fricatives (which are found in some Kriol/Light Warlpiri words of English origin). In the case of the Wubuy speakers, performance must indicate successful L2 learning, as Wubuy, like other Indigenous Australian languages, has no fricatives in its phonological inventory. None of the three participant groups were able to discriminate the voicing-based distinction between English /s z/ in word-initial position, a noteworthy failure of acquisition considering the very extended language contact situation in Australia, and the fact that, for the Kriol and Light Warlpiri speakers, this contact has given rise to new languages with (voiceless) fricative phonemes in the phonological inventories. We return to the /s z/ contrast in Section 4.2 below.

4.2. Acquisition of voicing distinctions in stop consonants

The results relevant to the question of acquisition of voicing as a new tier of contrastive information are less straightforward and consistent than the results for acquisition of frication as a MOA. Firstly, the studies indicate that speakers of Wubuy remain unable to discriminate voiced versus voiceless stops in any context, despite extended (lifelong) L2 English acquisition and use. This is perhaps surprising given the time-depth and quantity of exposure that the speakers have had to English. Kriol and Light Warlpiri speakers can discriminate voicing-based stop distinctions in bilabial stops, but they are not able to discriminate voicing-based stop contrasts at the velar POA. Kriol and Light Warlpiri speakers are (unsurprisingly) much better at discriminating English VOT-based stop contrasts when the voiceless stop is 'enhanced' to have a constriction duration that is consistent with a 'long stop' in Kriol and in Light Warlpiri, and in those traditional Indigenous Australian languages that have a fortis/lenis contrast in stops. As Wubuy does not implement duration-based contrasts in stops, this acoustic enhancement does not result in improved performance. As indicated in Section 4.1, none of the participant groups can discriminate the voicing-based fricative contrast /s z/.

4.3. Piggybacks: Ways into the inventory?

The group differences in discrimination success for fricative- and voicing-based contrasts, and the particular difficulty in discriminating voicing in fricatives, even for speakers who are lifelong L2 users, demand explanation. We speculate that, in cases where L1 perceptual attunement is particularly unhelpful because the crucial tier(s) of contrastive information falls outside of the 'language space', it may be the case that non-native contrasts are acquired by means of 'linguistic piggybacks'. These speculations rest on the assumptions that the human speech perception apparatus remains available across the lifespan and engages with any language acquisition task it encounters, using to its advantage (and sometimes disadvantage) skills and strategies developed in response to previous language experience. Where this language

experience comes up short in terms of solving a new language puzzle, we further assume that a learner will recruit and exploit any information, skill or strategy that offers itself as a key to identifying new language patterns (articulatory, or acoustic). In the following sections, we investigate the potential of 'linguistic piggybacks' to account for differences and similarities in the discrimination of voicing and fricative-based contrasts in the three studies included here.

4.3.1. Fricatives

The relative success in acquiring fricative MOA by the three participant groups may reflect at least two contributing factors, one perhaps classifiable as a piggyback (pertaining to the acquisition of frication as a MOA), and one as a more straightforward case of transfer from the L1 to the L2 (pertaining to the acquisition of different POAs in fricatives).

In terms of the acquisition of frication as a MOA, it is possible that non-speech use of /s/ (*sa!* is used in interjections to shoo dogs away in for instance Wubuy) provides sufficient experience to provide a piggyback for frication as a MOA. We may think of this as an analogous situation to what has been reported for the perception of Zulu click consonants by English speakers: clicks are often reported to be perceived as non-speech 'mouth sounds', used in relation to animals, yet English-acquiring infants and English-speaking adults [13] successfully discriminate Zulu click consonants. It is also possible that the three participant groups have other phonetic experiences with fricatives that may support the acquisition/discrimination of fricatives. In many Australian languages, stop consonants can be realized as approximants in certain contexts, resulting in phonetic fricatives at least some of the time [19]. Indeed, phonetic fricatives occurring in predictable environments provide some experience with fricatives as allophones—and may support the view that /s/ perhaps is not considered 'non speech' in the same manner as Zulu clicks, though [s] is unlikely to be one of the allophones of any of the native stops. Secondly, with respect to the acquisition of *place* distinctions in fricatives, we hypothesise that learners make good use of their L1 POA inventory, which may involve, as outlined in Section 3.1, up to six distinct POAs, including alveolar and palatal POAs. This means that while frication as a MOA may well be novel, the use of POA distinctions, once the manner is acquired, is a matter of little difficulty.

Existing research may shed additional light on the question of whether fricative POA contrasts are hard to acquire once the MOA has been noticed/acquired by a learner. For example, [12] found that English-learning 6- and 11-month-olds discriminate both non-native voiceless fricative POA contrasts (from non-native Nuu-Chah-Nulth and native English) with one within-organ and one between-organ contrast from each language. This result, though dealing with infants, suggests that non-native POA contrasts (or as the authors frame it, articulatory-organ differences) do not pose difficulties for infants. These results are somewhat parallel to our observations here, and they suggest that perception or acquisition of novel POA contrasts is scaffolded by existing linguistic experience (i.e. POA contrasts are implemented in consonant series with other MOAs) and perhaps also articulatory or other biases such as those proposed by AOH, or by models proposing Natural Referent Consonants [20]. In either case, it would appear that novel POA contrasts in a novel MOA (here, frication) pose little challenge to listeners: This is a tier of information that they have already mastered.

4.3.2. Voicing

The acquisition of voicing distinctions in stops and fricatives is a tale of some success but also of persistent difficulty for the participants in the studies reviewed here, and we again invoke the notion of 'linguistic piggybacks' in our account of the differences observed. Speakers of Wubuy, who do not implement an L1 voicing or duration contrast in stops, appear to fail to attend to VOT cues to contrast in both bilabial and velar stops. This is consistent with a 'use it or lose it'-principle, or at least with a need for extraordinary quality/quantity of L2 input to facilitate acquisition, beyond the lifelong exposure of the participants here. It would appear that these speakers are without L1 experience that allows them to 'unlock' voicing as a means to contrast maintenance: input from just a single series of stops does not provide necessary systematic phonetic variation for learners to use as a piggyback to noticing voicing as a potential means of contrast.

Speakers of the two contact varieties, Kriol and Light Warlpiri, each of which has stop duration contrasts (via substrate language influence) fare somewhat better, in particular with stop voicing contrasts with Kriol-like constriction durations. Speakers of Light Warlpiri additionally have some success with voicing only-contrasts. We suggest that L1 experience with long/short or fortis/lenis stops may provide the necessary piggyback for (at least partial) acquisition of contrastive voicing, as differences in proportion of VTT in long versus short stops may be systematic enough to feed a perceptual match with voiced versus voiceless English-derived stops. It is further the case that lenis stops have approximated allophones in languages with a duration contrast, but not the fortis stops [16]. Partial success aside, it remains, however, the case that 100+ years of continued language contact has not resulted in a 'full system' of contrast at all stop POAs. Finally, we speculate that velars may be difficult because they tend to be long, and the difference in proportion of voicing in long-short consonants may be less apparent.

5. Conclusions

The present paper discussed the question of whether an L2 learner can acquire an entire 'tier' of L2 contrasts that is not implemented in any form in their L1 language (L1). The studies reviewed suggest that the answer is 'Maybe': success may depend on the feature in question and a learner's linguistic experience. Fricative acquisition appears to be easier than acquisition of voicing, at least for the Australian sample here. The studies suggest that the strategy to success is using whatever systematic distributional information is available in the existing linguistic experience. This may be entirely coincidental (e.g., VTT) to the target tier, and the distributional information may not be sufficient (e.g., failure to acquire voicing in velars, fricatives). L2 learning is not always successful even when the learner is not directly constrained by conflict between the L1 and L2 phonological systems on shared tiers. It is possible that different organs/features/tiers may have different saliency to learners, related to biases in speech perception/production/processing, and in the extreme interpretation, this means that if you don't use it, you may lose it. One consequence of our findings is that some foundational features such as [±voice] may be lacking in languages such as those discussed here, providing evidence for an 'emergent' view of phonological features such as that espoused by [22].

6. Acknowledgements

We thank the participants, and the funding agencies: the Australian Research Council (DP170104457), the US National Science Foundation (#1348013), and the Australian National University (ANU) Futures Scheme grants.

7. References

[1] Best, C. T. (1994). The emergence of native-language phonological influences in infants: A perceptual assimilation model. *The development of speech perception: The transition from speech sounds to spoken words*, *167*(224), 233-277.

[2] Best, C. T. (1995). Learning to perceive the sound pattern of English. *Advances in infancy research*, *9*, 217-217.

[3] Best, C. T., & Tyler, M. D. (2007). Nonnative and second-language speech. *Language experience in second language speech learning: In honor of James Emil Flege*, *17*, 13.

[4] Jakobson, R., Fant, C. G., & Halle, M. (1951). *Preliminaries to speech analysis: The distinctive features and their correlates*. Cambridge, MA: MIT Press.

[5] Gibson, J. J. (1979) The ecological approach to visual perception. Boston: Houghton

[6] Browman, C. P., & Goldstein, L. (1990). Gestural specification using dynamically-defined articulatory structures. *Journal of Phonetics*, *18*(3), 299-320.

[7] Browman, C. P., & Goldstein, L. (1992). Articulatory phonology: An overview. *Phonetica*, *49*(3-4), 155-180.

[8] Browman, C. P., & Goldstein, L. (1995). Dynamics and articulatory phonology. *Mind as motion: Explorations in the dynamics of cognition*, *175*, 194.

[9] Browman, C. P., & Goldstein, L. (2000). Competing constraints on intergestural coordination and self-organization of phonological structures. *Les Cahiers de l'ICP. Bulletin de la communication parlée*, (5), 25-34.

[10] Goldstein, L., & Fowler, C. A. (2003). Articulatory phonology: A phonology for public language use. *Phonetics and phonology in language comprehension and production: Differences and similarities*, 159-207.

[11] Studdert-Kennedy, M., & Goldstein, L. (2003). Launching language: The gestural origin of discrete infinity. *Studies in the Evolution of Language*, *3*, 235-254.

[12] Tyler, M. D., Best, C. T., Goldstein, L. M., & Antoniou, M. (2014). Investigating the role of articulatory organs and perceptual assimilation of native and non-native fricative place contrasts. *Developmental psychobiology*, *56*(2), 210–227. https://doi.org/10.1002/dev.21195

[13] Best, C. T., McRoberts, G. W., & Sithole, N. M. (1988). Examination of perceptual reorganization for nonnative speech contrasts: Zulu click discrimination by English-speaking adults and infants. *Journal of experimental psychology. Human perception and performance*, *14*(3), 345–360. https://doi.org/10.1037//0096-1523.14.3.345

[14] Butcher, A. (2004). 'Fortis/lenis' revisited one more time: the aerodynamics of some oral stop contrasts in three continents. *Clinical linguistics & phonetics*, *18*(6-8), 547-557.

[15] Fletcher, J., & Butcher, A. (2014). Sound patterns of Australian languages. *The languages and linguistics of Australia: A comprehensive guide*, 91-138.

[16] Stoakes, H. M. (2013). *An acoustic and aerodynamic analysis of consonant articulation in Bininj Gun-wok* (Doctoral dissertation, University of Melbourne, School of Languages and Linguistics).

[17] Bundgaard-Nielsen, R. L., & Baker, B. J. (2016). Fact or furphy? The continuum in Kriol. *Loss and renewal: Australian languages since contact*, 177-216.

[18] Bundgaard-Nielsen, R. L., & O'Shannessy, C. (2021). When more is more: The mixed language Light Warlpiri amalgamates source language phonologies to form a near-maximal inventory. *Journal of Phonetics*, *85*, 101037.

[19] Ennever, T., Meakins, F., & Round, E. R. (2017). A replicable acoustic measure of lenition and the nature of variability in Gurindji stops. *Laboratory Phonology*, *8*(1).

[20] Best, C. T., Avesani, C., Tyler, M. D., & Vayra, M. (2019). PAM revisits the articulatory organ hypothesis : Italians' perception of English anterior and Nuu-Chah-Nulth posterior voiceless fricatives. In A. M. Nyvad, M. Hejná, A. Højen, A. B. Jespersen, & M. H. Sørensen (Eds.), A Sound Approach to Language Matters: In Honor of Ocke-Schwen Bohn (pp. 13-40). Retrieved from https://pure.au.dk/portal/files/153251381/Bohn_FINAL.pdf

[21] Bundgaard-Nielsen, R. L., Baker, B. J., Kroos, C. H., Harvey, M., & Best, C. T. (2015). Discrimination of multiple coronal stop contrasts in Wubuy (Australia): A natural referent consonant account. *PLoS One*, *10*(12), e0142054.

[22] Mielke, J. (2008). *The emergence of distinctive features*. Oxford University Press.

Adaptation to L3 Phonology? Perception of the Japanese Consonant Length Contrast by Learners of Italian

Kimiko Tsukada[1, 2], *John Hajek*[2]

[1]Macquarie University, [2]The University of Melbourne
kimiko.tsukada@gmail.com, j.hajek@unimelb.edu.au

Abstract

Both Italian and Japanese use consonant length contrastively, but length contrast is known to be difficult for non-native speakers. The perception of Japanese singleton/geminate contrasts by learners of Italian from American English and Argentinian Spanish backgrounds was compared to determine if learning Italian may be helpful for processing known contrasts in an unknown language. The two groups did not differ significantly and both groups discriminated length contrasts more accurately when alveolar geminate [tː] (rather than velar singleton [k]) occurred in the target position. The results suggest that knowledge of Italian may not automatically guarantee adaptation to Japanese length processing.

Index Terms: Japanese consonant length, American English, Argentinian Spanish, Italian, second/foreign language (L2/FL)

1. Introduction

Although typologically unrelated, both Italian and Japanese use consonant length contrastively [1-3]. For example, in Italian, *eco* and *ecco* mean 'echo' and 'here (it is)', respectively. In Japanese, on the other hand, *ika* 以下 and *ikka* 一課 mean 'below' and 'lesson one', respectively. This differs from languages such as English or Spanish, neither of which has an underlying consonant length contrast at the level of words [4, 5] (with a possible exception of the tap vs trill contrast in Spanish [6]). Length contrasts are important for communication purposes as shown above, but they are known to be difficult for non-native speakers from diverse L1 backgrounds [3, 7, 8].

The question of interest in this study is whether individuals who have been exposed to Italian as opposed to Japanese consonant length may adapt themselves efficiently to Japanese consonant length. This can be characterized as the indirect influence of L2/FL (second/foreign language), rather than L1, learning on the processing/acquisition of L3 (third language) phonology. This is a novel aspect of our study, as there is still little research [9, 10] that investigates the transfer effect of L2/FL learning on L3 phonology across typologically unrelated languages such as Japanese and Italian. There is general agreement that L2 plays an important role in the L3 acquisition process, especially at the initial stage of learning [11, 12].

Japanese is a popular L2/FL in many countries including Australia and the USA. In addition to local native English-speaking learners, there are many learners from non-English-speaking backgrounds (e.g., Mandarin Chinese or Korean) in large cities such as Sydney or Melbourne in Australia. Typically, Japanese is their L3 after L2 English for these learners. It is therefore important to gain a better understanding of the processing/acquisition of difficult Japanese sounds by non-native learners from diverse linguistic backgrounds and utilize the knowledge for improving pronunciation pedagogy.

In this study, we gained access to and compared two groups of participants who had exposure to Italian but differed in their L1 (American English vs Argentinian Spanish). As neither English nor Spanish uses consonant length contrastively, it may be their shared L2/FL, Italian, rather than their respective L1s that may influence their perception of unfamiliar Japanese singleton/geminate. While L2/FL Italian learning experience may be expected to enhance consonant length perception in general, it is possible that the benefit of learning may be specific to the processing of the speech sounds of the target language, Italian. This is because some cross-linguistic differences between Italian and Japanese in consonant length have been reported. For instance, vowels preceding geminates are shorter than vowels preceding singletons in Italian [13], but longer in Japanese [14, 15]. Another example is that liquid geminates occur frequently in Italian, but not in Japanese [14].

In fact, the rate of gemination in Italian written texts has also been reported to depend on the phoneme type such that long /tː/ occurs at only 15% of the rate of short /t/, whereas long /kː/ is even less frequent at a rate of only 5% of all cases of short /k/, respectively [Table 3 in 16]. This appears to differ from the frequency of occurrence of Japanese geminates. According to [17], long /kː/ occurs at a rate of 32% of all geminate consonants, whereas long /tː/ at a rate of 28% in the corpus of a Japanese newspaper. Geminate /k/ was also more frequent than geminate /t/ in the Corpus of Spontaneous Japanese and the ratio of singleton/geminate closure duration was larger for /k/ (2.84) than for /t/ (2.38) [14]. If so, short/singleton vs long/geminate /k/ may be acoustically more salient than other contrasts in Japanese. These cross-linguistic phonetic differences may affect how learners of Italian perceive Japanese singleton/geminate consonants and we were also interested in determining if listeners' length perception accuracy varies depending on the type of consonant by comparing alveolar and velar places of articulation.

Unfortunately, at present, we do not know how our participants perceive the Italian singleton/geminate contrast, which is a serious limitation of the present study. However, the accepted wisdom is that learning to communicate efficiently in an FL is a challenge for most adults and requires a huge investment of time and effort. Therefore, we were motivated to study if and how, not only learners of Japanese, but learners of Italian may indirectly benefit from FL experience by examining known singleton/geminate contrasts in an unknown language, Japanese. We also present data collected from American English learners of Japanese for comparison.

2. Method

The experimental stimuli and procedures were identical to those used previously [18].

2.1. Stimuli preparation

2.1.1. Speakers

Six (3 males, 3 females) native speakers of Japanese participated in the recording sessions, which lasted between 45 and 60 minutes. The speakers' age ranged from late twenties to early forties. All speakers spoke standard Japanese, having been born or having spent most of their life in the Kanto region surrounding the Greater Tokyo Area [20 for references to cross-dialectal studies]. The first author (native speaker of Japanese originally from Tokyo) auditorily confirmed that all the speakers clearly differentiated the singleton and geminate consonants by duration. The speakers were recorded in the recording studio at the National Institute of Japanese Language and Linguistics, Tokyo.

2.1.2. Speech materials

Table 1 shows 12 Japanese word pairs used in this study. The /(C)VC(C)V/ tokens contained singleton (n = 96) or geminate (n = 96) consonants intervocalically (underlined). Only tokens with stops were considered in this study. As voiced geminates are disfavoured and their occurrence is limited in Japanese [14, 19, 20], only voiceless stops (/t, k/) were used. On average, the closure durations were 96 ms and 262 ms for singletons and geminates, respectively. Averaged across tokens by all speakers, the geminate-to-singleton ratios were 2.7 for alveolars (/t/-/t:/) and 2.8 for velars (/k/-/k:/), respectively. These durational values are in good agreement with what has been reported in previous research [21] (see, however, [14] for alveolars).

Table 1. *Twelve pairs of Japanese words used with target sounds underlined and bolded. HL and LH indicate High-Low and LH pitch patterns, respectively.*

	Singleton		**Geminate**	
/t/	*he**t**a*^{LH}	'unskilled'	*he**tt**a*^{LH}	'decreased'
	*ka**t**o*^{HL}	'transient'	*ka**tt**o*^{HL}	'cut'
	*ma**t**e*^{HL}	'wait'	*ma**tt**e*^{HL}	'waiting'
	*o**t**o*^{LH}	'sound'	*o**tt**o*^{LH}	'husband'
	*sa**t**e*^{HL}	'well, then'	*sa**tt**e*^{HL}	'leaving'
	*wa**t**a*^{LH}	'cotton'	*wa**tt**a*^{LH}	'broke'
/k/	*a**k**e*^{LH}	'open'	*a**kk**e*^{LH}	'appalled'
	*ha**k**a*^{LH}	'grave'	*ha**kk**a*^{LH}	'mint'
	*i**k**a*^{HL}	'below'	*i**kk**a*^{HL}	'lesson one'
	*ka**k**o*^{HL}	'past'	*ka**kk**o*^{HL}	'parenthesis'
	*sa**k**a*^{LH}	'slope'	*sa**kk**a*^{LH}	'author'
	*shi**k**e*^{LH}	'rough sea'	*shi**kk**e*^{LH}	'humidity'

2.2. Participants

Five (one native and four non-native) groups of participants took part in the AXB discrimination task. Our target groups are learners of Italian whose L1 was either American English (AE) or Argentinian Spanish (AS). The former group (AE + Italian) consisted of 7 (3 males, 4 females, *mean age* = 25.3 years, *sd* = 11.8) participants, started learning Italian at the age of 18.3 on average (*sd* = 2.0) and had a mean length of learning of 2.6 (*sd* = 2.2) years. Three of them were enrolled in the first-year, another three in the second-year, one each in the third-year and fourth-year level Italian at University of Oregon in Eugene, OR, USA. The latter group (AS + Italian) also consisted of 7 (2 males, 5 females, *mean age* = 24.0 years, *sd* = 5.8) participants. Three were in the second-year, three were in the third-year and one was at the fourth-year level. They were enrolled in the Universidad Nacional del Litoral in Sante Fe, Argentina. As the number of participants in both groups is small, the results need to be regarded as preliminary.

Two other non-native groups consisted of AE speakers who were students at University of Oregon. Some of their results were reported previously [22] and included here only for comparison. One group consisted of 19 (7 males, 12 females, *mean age* = 22.9 years, *sd* = 3.7) AE learners of Japanese (AE + Japanese) at different levels of proficiency. Two of them were heritage learners. One was enrolled in the first-year level and the other was enrolled in the third-year level Japanese. Excluding these heritage learners, the AE + Japanese listeners started learning Japanese at the age of 17.4 on average (*sd* = 5.8) and had a mean length of learning of 4.5 (*sd* = 4.1) years. The other group consisted of 17 (4 males, 13 females, *mean age* = 19.8 years, *sd* = 1.0) AE speakers inexperienced in Japanese who were enrolled in Psychology or Linguistics courses and received credit for research participation. Neither of these two AE groups had experience learning Italian formally at college level, but they differed in their experience with Japanese. However, this is not intended to guarantee that the two groups are comparable apart from Japanese experience and participants' language background needs to be more tightly controlled in future work.

The last and a control group consisted of 10 (2 males, 8 females) native speakers of Japanese (NJ) who were students at University of Oregon. All NJ speakers were born and spent the majority of their life in Japan. Their mean length of residence in the US was 0.4 years (*sd* = 0.22) at the time of participation. None of the NJ speakers participated in the recording sessions. According to self-report, all five groups of participants had normal hearing.

All participants were tested individually in a session lasting approximately 30 to 40 minutes in a sound-attenuated laboratory or a quiet room at their own university. The experimental session was self-paced. The participants heard the stimuli at a self-selected, comfortable amplitude level over the high-quality headphones on a computer.

2.3. Procedure

The participants 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 [23]. 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., 'yoka$_2$'-'yoka$_1$'-'yokka$_3$') or B (e.g., 'soto$_3$'-'sotto$_1$'-'sotto$_2$'; where the subscripts indicate different speakers).

The participants listened to a total of 200 trials. The first eight trials were for practice and were not analyzed. 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 group 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 [24]. All possible AB combinations (i.e., AAB, ABB, BAA, and BBA, 48 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. They were informed that they would hear words from languages which may be unfamiliar to them, but they were not explicitly instructed to listen for the difference in consonant length. No feedback was provided during the experimental sessions. The participants could take a break after 50 trials if they wished (information on this is available, but not analyzed). 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 3.6.0 for preliminary statistical analyses and data visualization reported below [25]. The packages used include ez [26] and tidyverse [27].

3.1. Overall results

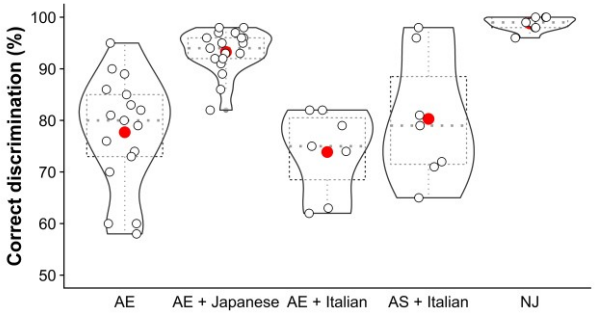


Figure 1: *The distributions of length discrimination accuracy (%) by five groups of participants. The horizontal line and the red circle in each box indicate the median and mean, respectively.*

Figure 1 shows the distributions of percentages of correct discrimination by the five groups of participants. The overall mean discrimination accuracy for non-native participants was 78%, 93%, 74% and 80% for the AE, AE + Japanese, AE + Italian and AS + Italian groups, respectively. The NJ group was at near ceiling (99%) with little individual variation. It is clearly visible that the AE + Japanese group with direct learning experience outperformed the other non-native groups. For AE listeners (i.e., AE vs AE + Italian), it appears that learning Italian with singleton/geminate contrasts did not positively transfer to the perception of consonant length in unfamiliar Japanese. A direct comparison of our two target groups of learners of Italian, AE + Italian and AS + Italian, via the Welch two-sample *t*-test showed that the between-group difference was not significant [$t(10.43) = -1.1$, $p = 0.29$].

3.2. Comparison of the direction of category change (Geminate > Singleton or Singleton > Geminate)

Next, we compared the results for the trials in which the consonant length in the AXB sequence changed from 1) geminate (G) to singleton (S) (i.e., G-G-S, G-S-S) or 2) S to G (i.e., S-S-G, S-G-G), focusing on the two target groups. This analysis would enable us to evaluate if the participants detected a change in length category bi-directionally or not. Figure 2 shows the distributions of percentages of correct discrimination by the two groups of learners of Italian as a function of the direction of category change within a trial (G > S, S > G). The AE + Italian group showed nearly identical discrimination accuracy whether the direction of change was from G to S (74%) or from S to G (73%). The AS + Italian group, on the other hand, was slightly more accurate when the direction of change was from S to G (83%) than when it was from G to S (79%). Two-way analysis of variance (ANOVA) with group (AE + Italian, AS + Italian) as a between-subjects factor and direction of category change (G > S, S > G) as a within-subjects factor did not reach significance for the main effects nor the interaction effect.

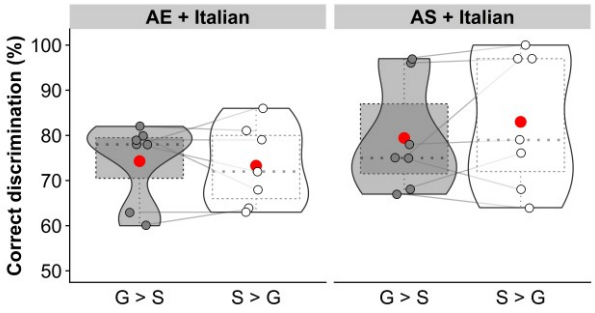


Figure 2: *The distributions of length discrimination accuracy (%) by two groups of learners of Italian for trials differing in the direction of category change (Geminate > Singleton, Singleton > Geminate). The light lines connect individual participants' scores.*

3.3. Comparison of the length category (Geminate vs Singleton) of the target token (X in AXB)

In the AXB discrimination task, the identity of the token placed in the target position may affect the participants' discrimination accuracy. Figure 3 shows the distributions of percentages of correct discrimination by the two groups of learners of Italian as a function of the length category of the target token (geminate vs singleton).

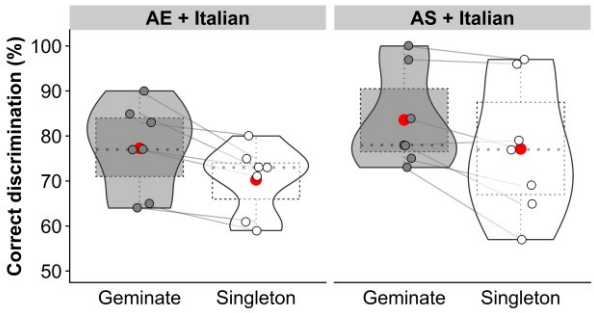


Figure 3: *The distributions of length discrimination accuracy (%) by two groups of learners of Italian for trials differing in the length category of the target token (Geminate, Singleton).*

When the length category of the target token (i.e., X in the AXB sequence) was taken into consideration, both groups were more accurate when the target consonant was a geminate than when it was a singleton (AE + Italian: 77% vs 70%, AS + Italian: 84% vs 77%). Two-way ANOVA with group (AE + Italian, AS + Italian) as a between-subjects factor and length (geminate, singleton) as a within-subjects factor reached

significance only for the main effect of length [$F(1, 12) = 20.8$, $p < 0.001$, $\eta_G^2 = .09$]. The lack of significant interaction suggests that both groups discriminated length contrasts more accurately when the target consonant was a geminate than when it was a singleton.

3.4. Comparison of length discrimination at alveolar (/t/-/tː/) and velar (/k/-/kː/) places of articulation

As mentioned in the Introduction, the relative frequency of consonants occurring as geminates appears to differ between Italian and Japanese. Of relevance to the present study, /tː/ occurs more frequently than /kː/ in Italian, but the reverse seems to be the case in Japanese. Given this cross-linguistic difference in distributional frequency, in this section, we provide an analysis of whether or not the place of articulation (i.e., alveolar vs velar) may affect listeners' length discrimination accuracy. Figure 4 shows the distributions of percentages of correct length discrimination by the two groups of learners of Italian and a group of learners of Japanese as a function of the place of articulation (alveolar vs velar) and the length category (geminate, singleton) of the target token.

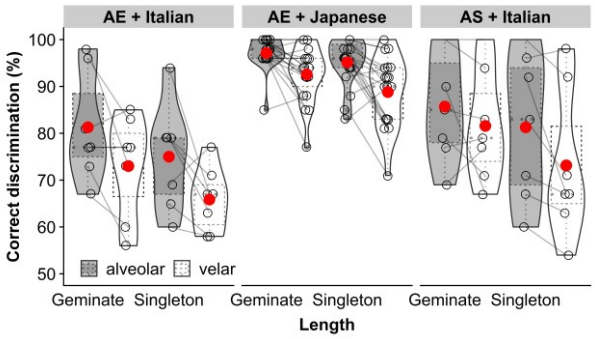


Figure 4: *The distributions of length discrimination accuracy (%) by two groups of learners of Italian and a group of learners of Japanese for trials differing in the place of articulation of the target token (alveolar, velar).*

When the place of articulation (alveolar vs velar) of the target token was taken into consideration, all three groups more accurately discriminated consonant length when the target consonant was alveolar than when it was velar (AE + Italian: 78% vs 69%, AE + Japanese: 96% vs 91%, AS + Italian: 83% vs 77%). This held true for both length categories as shown in Figure 4 (AE + Italian: 81% vs 73% for geminate, 75% vs 66% for singleton, AE + Japanese: 97% vs 93% for geminate, 95% vs 89% for singleton, AS + Italian: 86% vs 82% for geminate, 81% vs 73% for singleton). Three-way ANOVA with group (AE + Italian, AE + Japanese, AS + Italian) as a between-subjects factor and length (geminate, singleton) and place (alveolar, velar) as within-subjects factors reached significance for the main effects of group [$F(2, 30) = 20.0$, $p < 0.001$, $\eta_G^2 = .49$], length [$F(1, 30) = 29.0$, $p < 0.001$, $\eta_G^2 = .06$] and place [$F(1, 30) = 50.9$, $p < 0.001$, $\eta_G^2 = .12$] only. No interactions were significant. The lack of significant interactions suggest that the effects of length and place were comparable for all three groups despite different L2/FL learning backgrounds.

4. Discussion

This study examined the perception of Japanese consonant length contrasts by learners of Italian whose L1 was American English (AE + Italian) or Argentinian Spanish (AS + Italian). We were interested in determining whether individuals who were exposed to Italian consonant length may adapt themselves efficiently to consonant length in an unfamiliar language. As the number of participants in each group is small, the data are still preliminary. The two groups of learners of Italian did not differ significantly from each other and were much less accurate than native and non-native speakers of Japanese (Figures 1 and 4). Further, both AE + Italian and AS + Italian (and AE + Japanese) groups discriminated length contrasts more accurately when a geminate or alveolar stop rather than a singleton or velar stop occupied the target (X) position in the AXB sequence. This resulted in a hierarchy of discriminability with [tː] at the top and [k] at the bottom (Figure 4).

The participants' lack of adaptation to Japanese singleton/geminate may be related to cross-linguistic differences (e.g., duration of adjacent vowels, etc.) in how Italian and Japanese consonant length contrasts are realized. It is difficult to hypothesize that the place effect observed in this study is related to the frequency distribution of Italian consonants that occur as geminates [16]. This is because AE + Japanese who were inexperienced in Italian showed the same pattern of results as the two groups of learners of Italian (Figure 4). While the stimuli used in the present study had almost identical geminate-to-singleton ratios for both stops (2.7 for alveolars and 2.8 for velars, respectively), there may be some acoustic phonetic characteristics other than durational ratios that make the velar consonant hard to perceive as pointed out by one reviewer. In this connection, previous research reported that the geminate /t/ was less accurately perceived than the geminate /s/ by native speakers of German despite larger geminate-to-singleton ratios for the former than the latter [28].

What is crucially lacking in this study is how these participants perceive Italian singletons and geminates. In future work, it is necessary to examine participants' consonant length processing in both familiar and unfamiliar languages to assess if and how consonant length perception is related in known and unknown languages. It is also necessary to increase the number of participants and further investigate how the proficiency of Italian may affect the perception of both Italian and Japanese consonant length contrasts. Unlike Italian, both vowel as well as consonant length is contrastive in Japanese and both short and long vowels can precede singletons and geminates. It would be interesting to examine how phonological vowel length (as opposed to phonetic vowel lengthening) may influence the perception of the length category of the following consonants.

5. Conclusions

The learners of Italian from American English and Argentinian Spanish backgrounds did not differ from each other in discriminating consonant length contrasts in Japanese. This may be because they lack specific knowledge of the phonetic characteristics of Japanese singletons and geminates. Both groups also discriminated length contrasts more accurately when alveolar geminate [tː] (rather than velar singleton [k]) occurred in the target position. The results suggest that experience with Italian singletons and geminates may not automatically transfer to the efficient processing of Japanese singletons and geminates.

6. Acknowledgements

This work is supported by the 2018 Endeavour Research Fellowship and the Sumitomo Foundation Fiscal 2020 Grant for Japan-related Research Projects awarded to the first author. We

thank Kaori Idemaru for the use of Spoken Language Research Laboratories and participants for making the study possible. We also thank three anonymous reviewers for their constructive comments and suggestions for improving this work.

7. References

[1] Vance, T. J., The Sounds of Japanese. Cambridge: Cambridge University Press, 2008.
[2] Rogers, D. and d'Arcangeli, L. "Italian", Journal of the International Phonetic Association, 34(1): 117-121, 2004.
[3] De Clercq, B., Simon, E. and C. Crocco, C., "Rosa versus rossa: The acquisition of Italian geminates by native speakers of Dutch", Phrasis: Studies in Language and Literature, 2: 3-29, 2014.
[4] Kaye, A. S., "Gemination in English", English Today, 21: 43-55. 2005.
[5] Oh, G. E. and Redford, M. A., "The production and phonetic representation of fake geminates in English", Journal of Phonetics, 40: 82-91, 2012.
[6] Cordero, D., Ruiz-Peña, E., Sierra, E., Stevenson, R. and Rafat, Y., "Second dialect and second language imitation of geminates by Colombian Spanish speakers", in E. Babatsouli [Ed], Proceedings of the International Symposium on Monolingual and Bilingual Speech 2017, 99-105, 2017.
[7] Han, M. S., "The timing control of geminate and single stop consonants in Japanese: A challenge for non-native speakers", Phonetica, 49: 102-127, 1992.
[8] Kubozono, H., "Introduction to the special issue on Japanese geminate obstruents", Journal of East Asian Linguistics, 22: 303-306, 2013.
[9] Cabrelli Amaro, J. and Wrembel, M., "Investigating the acquisition of phonology in a third language: A state of the science and an outlook for the future", International Journal of Multilingualism, 13(4): 395-409, 2016.
[10] Magdalena, W., Marecka, M. and Kopečková, R., "Extending perceptual assimilation model to L3 phonological acquisition", International Journal of Multilingualism, 16(4): 513-533, 2019.
[11] Gut, U., "Cross-linguistic influence in L3 phonological acquisition", International Journal of Multilingualism, 7(1): 19-38, 2010.
[12] Wrembel M., "L2-accented speech in L3 production", International Journal of Multilingualism, 7(1): 75-90, 2010.
[13] Hajek, J., Stevens, M. and Webster, G. "Vowel duration, compression and lengthening in stressed syllables in Italian", Proceedings of the 16th ICPhS, 1057-1060, 2007.
[14] Sano, S., "The distribution of singleton/geminate consonants in spoken Japanese and its relation to preceding/following vowels", Proceedings of 19th ICPhS, 1833-1837, 2019.
[15] Idemaru, K. and Guion-Anderson, S., "Relational timing in the production and perception of Japanese singleton and geminate stops", Phonetica, 67, 25-46, 2010.
[16] Arango, J., DeCaprio, A., Baik, S., De Nardis, L., Shattuck-Hufnagel, S. and Di Benedetto, M.-G., "Estimation of the frequency of occurrence of Italian phonemes in text", Online: https://arxiv.org/pdf/2101.06147.pdf (accessed on June 7, 2022).
[17] Tamaoka, K. and Makioka, S., "Frequency of occurrence for units of phonemes, morae, and syllables appearing in a lexical corpus of a Japanese newspaper", Behavior Research Methods, Instruments, & Computers, 36(3): 531-547, 2004.
[18] Tsukada, K., Yurong, Kim, J.-Y., Han, J.-H. and Hajek, J., "Cross-linguistic perception of the Japanese singleton/geminate contrast: Korean, Mandarin and Mongolian compared", Proceedings of 22nd Interspeech, 3910-3914, 2021.
[19] Hussein, Q. and Shinohara, S., "Partial devoicing of voiced geminate stops in Tokyo Japanese", The Journal of the Acoustical Society of America, 145: 149-163, 2019.
[20] Kawahara, S., "The phonetics of sokuon, or geminate obstruents", in H. Kubozono [Ed], Handbook of Japanese Phonetics and Phonology, 43-78, Berlin: Walter de Gruyter, 2015.
[21] Hayes-Harb, R., "Optimal L2 speech perception: Native speakers of English and Japanese consonant length contrasts", Journal of Language and Linguistics, 4: 1-29, 2005.
[22] Tsukada, K., Idemaru, K. and Hajek, J., "The effects of foreign language learning on the perception of Japanese consonant length contrasts", Proceedings of 17th SST, 37-40, 2018.
[23] Boersma, P. and Weenink, D., Praat: Doing Phonetics by Computer [version 6.0.19], retrieved from http://www.praat.org (Last viewed June 13, 2016).
[24] 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.
[25] R Core Team, R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/ 2019.
[26] Lawrence, M. A. ez: Easy Analysis and Visualization of Factorial Experiments. R package version 4.4-0. https://CRAN.R-project.org/package=ez 2016.
[27] 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.
[28] Oba, R., Braun, A. and Handke, J., "The perception of Japanese geminates by native and non-native listeners", The Phonetician, 92: 9-29, 2005.

Stop (de)gemination in Veneto Italian: The role of durational correlates

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 investigates a long-assumed but previously untested degemination of stops in the regional variety of Italian spoken in the Veneto, in North-East Italy. The durational parameters known to be affected by gemination in Italian – i.e., consonant duration, duration of the preceding vowel and the ratio between the two – are considered. The entire Italian stop series is investigated through an acoustic-phonetic production experiment involving six speakers reading a set of carrier sentences designed to elicit different prosodic patterns. Partial degemination is observed for most speakers in terms of (a) decreased geminate-singleton consonant duration differences compared to previous studies on other Italian varieties, and (b) considerable overlap between geminate and singleton consonant-to-vowel duration ratio categories. Possible sociophonetic effects are discussed.

Index Terms: gemination, degemination, duration ratio, regional Italian, Veneto Italian.

1. Introduction

1.1. Background

1.1.1. Stop gemination and its primary acoustic correlates in Italian

Italian displays a phonological contrast word-medially between long (or geminate) and short (or singleton) consonants. This length contrast involves the entire Italian stop series /p p: t t: k k: b b: d d: g g:/, as in *fato* /fato/ 'fate' vs *fatto* /fat:o/ 'fact' [1]-[3]. Gemination is realized phonetically primarily by increased consonant (C) duration and decreased duration of the preceding (stressed) vowel (V1) for geminates in this language [4]-[6]. Italian vowels are predictably long in word-medial position when the stressed syllable is open, e.g., in [ˈfa:.to], and short in the same position when the stressed syllable is closed, e.g., [ˈfat.to]. Vowels in unstressed syllables, on the other hand, are short regardless of whether the syllable is open (e.g., in *fatina* [fa.ˈti:.na] 'fairy') or closed (e.g., *fattore* [fat.ˈto:.re] 'factor') [14]. Another proposed acoustic correlate of Italian gemination is the ratio between the duration of the consonant and that of the preceding vowel (C/V1) [7]. Variation in C/V1 has been claimed to be a more stable correlate of gemination across speaking rates than variation in absolute C duration values. In particular, for Italian post-stress stops a C/V1 ratio lower than 1.00 would indicate a singleton and the same ratio higher than 1.00 a geminate; for pre-stress stops, on the other hand, the cut-off ratio would be 2.00 [7].

The available empirical knowledge about Italian gemination derives from previous laboratory-based studies mostly concerning the Tuscan and other Central-Southern (mainly Roman) regional varieties of Italian which are historically strongly associated with Standard Italian (see § 1.1.2). In terms of phonetic C duration, the previous findings point towards a clear-cut distinction between phonological length types, with geminate stops found to be roughly twice as long as singletons in nuclear accented position within an intonational phrase [3], [6]-[10]. For V1 duration, on the other hand, substantial inter-speaker variation in the magnitude of stressed pre-geminate vowel shortening (roughly between 20 and 50% of total V1 duration) has been reported across a number of studies (e.g., [4]-[7], [11]-[14]), suggesting that this phenomenon may be more variable in nature. It has also been proposed that V1 duration may be directly linked with the phonetic duration of C rather than its phonological length, with C and stressed V1 duration appearing to be in an inverse linear relationship [15]. This would explain the less 'categorical' gemination-triggered behavior of V1 duration as opposed to that of C duration reported in the literature.

1.1.2. Cross-regional differences and Northern degemination

It is important to note that Italian is not a homogeneous language at the phonetic level, particularly from a geographic perspective. Indeed, notable cross-regional differences exist between Northern varieties on the one hand, and Central-Southern varieties on the other in this respect [1]. The Central-Southern varieties are said to comply with Standard Italian pronunciation norms in terms of consonant length, with clear duration differences between geminates and singletons reported in the numerous studies focusing on these varieties. On the other hand, the Italian spoken in Northern Italy, and particularly in the Veneto Region in North-East Italy (henceforth, Veneto Italian), has often been associated with consonant degemination [16], [17]. According to these claims, Northern Italian geminates are generally produced with shorter duration [18] or virtually no duration difference from singletons [17]. However, these claims stem from impressionistic auditory descriptions and as such have mostly not been backed by experimental acoustic evidence (cf. [19] for a discussion). Indeed, in the only large-scale, cross-regional study on Italian gemination [19] no significant geminate-to-singleton C duration differences between Northern and Central-Southern varieties of Italian were observed.

1.1.3. Veneto Italian stop (de)gemination

To our knowledge, no relevant empirical acoustic information on stop gemination can be extracted from studies specifically addressing Veneto Italian, as this variety has been particularly understudied in terms of experimental phonetic research. This is striking considering the claims strongly associating this variety with degemination, e.g., [16]-[18]. These claims may stem from the fact that many Veneto speakers are likely to show a relatively strong influence of the locally spoken Romance dialect (which lacks a consonant quantity distinction) on their Italian, as regional dialect is more widely spoken in the Veneto compared to other areas in Italy [20]. Indeed, many Veneto speakers are said to be bilectal, i.e., actively speaking both Italian and the dialect to different degrees [20].

1.2. Aims

In the first instance this study aims at providing some experimental durational acoustic findings on stop gemination for Veneto Italian, a regional variety for which they are lacking. Another aim is to test whether the claims regarding degemination hold true for this variety. This is done by considering the established durational acoustic cues to gemination for Italian (C, V1 duration) and their ratio (C/V1).

2. Methods

2.1. Participants

The participants were three females and three males all born, raised, and living in the south-western part of the province of Vicenza, in central Veneto, at the time of the experiment. They were all active speakers of the local dialect. Unusually, VenS3F had studied Standard Italian elocution not long before the experiment. The list of participants with relevant sociophonetic information is listed in Table 1. The 'Relative Italian-dialect bilectalism' column provides information as to whether the dialect or Italian was predominant, or whether the participants made a balanced used of both lects.

Table 1. *List of participants with related sociophonetic information.*

Participant ID	Sex	Age	Level of Schooling	Rel. Italian-dialect bilectalism
VenS1F	F	68	Elem. school	Dialect
VenS2F	F	43	Sec. school	Balanced
VenS3F	F	44	Uni. degree	Balanced
VenS1M	M	50	Uni. degree	Balanced
VenS2M	M	34	Sec. school	Balanced
VenS3M	M	35	Sec. school	Balanced

2.2. Material and procedure

An acoustic phonetic experiment was set up to collect the data, namely C and V1 duration values. C covered all Italian stop phonemes while V1 included only mid or low vowels. The entire experiment was conducted in Italian. The phones of interest were embedded in real Italian target words forming part of read carrier sentences, with the words either in sentence-final (*ho detto WORD*, 'I said WORD') or sentence-medial (*ho detto WORD prima*, 'I said WORD before') position. The participants were asked to read the carrier sentences as if they were answering a question which placed the focus on the target word (*Che hai detto?/Che hai detto prima?*, 'What did you say?/What did you say before?). Both positions were designed to elicit nuclear accented target words. The target words also had two different stress conditions (post-stress or pre-stress) based on the position of the target consonant relative to that of lexical stress. The words were all paroxytones; post-stress words were disyllabic and pre-stress words trisyllabic. This means that V1 was stressed in post-stress words and unstressed in pre-stress words. The resulting twenty-four target words are listed in detail in Table 2 below. Target /V1-C(C)/ sequences are in bold.

The carrier sentences were presented on a computer screen in random order and repeated by each of the six speakers four times, for each of the twenty-four target words, and for each of the two position-in-the-phrase conditions. The total number of tokens was therefore 1152. The sentences were recorded through a professional solid-state recorder, sampled at 44100 Hz with 16-bit quantization.

Table 2. *List of experimental words.*

Phon	Post-stress		Pre-stress	
	Sing	Gem	Sing	Gem
/p/	*Papa* /ˈ**pa**pa/ 'Pope'	*pappa* /ˈ**pap**pa/ 'mush'	*Papato* /pa**ˈp**ato/ 'Papacy'	*pappato* /p**ap**ˈpato/ 'gobbled up'
/t/	*fato* /ˈ**fa**to/ 'fate'	*fatto* /ˈ**fat**to/ 'fact'	*patata* /pa**ˈt**ata/ 'potato'	*dettato* /d**et**ˈtato/ 'dictation'
/k/	*paca* /ˈ**pa**ka/ 'he/she calms'	*pacca* /ˈ**pak**ka/ 'pat'	*pacato* /pa**ˈk**ato/ 'placid'	*paccato* /p**ak**ˈkato/ 'let down' (p.p.)
/b/	*roba* /ˈ**rɔ**ba/ 'stuff'	*gobba* /ˈ**gɔb**ba/ 'hump'	*Babele* /ba**ˈb**ɛle/ 'Babel'	*dabbene* /d**ab**ˈ**b**ɛne/ 'respectable'
/d/	*cade* /ˈ**ka**de/ 'he/she falls'	*cadde* /ˈ**kad**de/ 'he/she fell'	*badato* /ba**ˈd**ato/ 'looked after'	*laddove* /l**ad**ˈ**d**ove/ 'whereby'
/g/	*lega* /ˈ**le**ga/ 'he/she ties'	*legga* /ˈ**lɛg**ga/ 'read' (pr.sub.)	*pagato* /pa**ˈg**ato/ 'paid'	*taggato* /t**ag**ˈ**g**ato/ 'tagged'

2.3. Analysis

The phonetic annotation of the target words was done manually in EMU-SDMS [21] and the duration of the target segments was extracted and analyzed through emuR [22]. For voiceless stops, V1 onset and offset were placed at the start and end of V1, respectively – i.e., where periodicity and glottal pulses were visible in the waveform and spectrogram. C onset coincided with V1 offset and C offset with the onset of phonation in the following vowel. Thus, the release phase (indicated as [h] in Figure 1) was included in C duration (as in e.g. [3], [9]). Voiced stop onset and offset boundaries were placed where a sudden (a) drop and (b) rise of amplitude and F2 energy were visible in the waveform and spectrogram, respectively [3]. A release burst, where present, was included in C duration. An example of annotation for singleton /fato/ is shown in Figure 1 below.

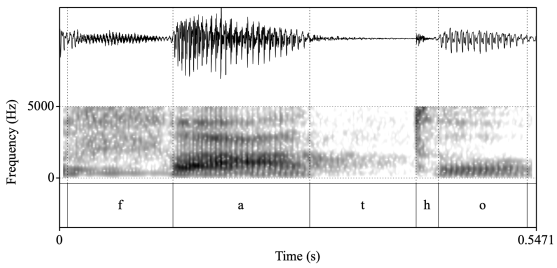


Figure 1: *Example of annotation of token /fato/ uttered by VenS3F. V1* = [a]; *C* = [t^h].

Average C and V1 duration values, and C/V1 duration ratios were calculated for each length type and each speaker, with the aim of observing differences between length types. To

attest the statistical significance of the results, quantitative data analyses on the raw data were performed in R through Linear Mixed-Effects Models using the lmerTest package [23]. Three models were run – one for each of the dependent variables: i) C duration; ii) V1 duration; and iii) C/V1 ratio. Each of the three dependent variables was modeled as a function of gemination and other fixed-effects factors, namely voicing, stress condition, position in the phrase, place of articulation, and speaker sex, as well as their interactions. The models where initially maximally specified and subsequently step-reduced (through the step function that is part of [23]) to obtain the best-fitting models. The participants were treated as a random-effect factor with by-speaker varying intercepts and varying gemination slopes. The models found significant effects for various factors; however, due to the limited space available, only the results for gemination and relevant interactions are reported. The emmeans function (part of [24]) was employed to conduct post-hoc tests so as to investigate differences between levels in interactions.

3. Results

3.1. C duration

The results for C duration are summarized for each speaker in Table 3 below. They support the partial degemination claims, although not for all speakers. Overall, the geminate-to-singleton (CC/C) duration ratio is 1.76, which is lower than the ratio of 2.00 reported in the literature for similar experimental settings concerning the central varieties. However, VenS3F – the speaker who had previously studied Standard Italian elocution – produced a ratio greater than 2.00.

The statistical analysis confirmed the presence of a significant main effect of gemination on the mixed-effects model across speakers and conditions (β = -92.90 ms, SE = 13.01 ms, t = -7.14, p < .001).

Table 3. *Mean duration (in ms), SD (in parentheses), and counts (n) of geminate and singleton C for each speaker across conditions. CC/C ratios are also provided.*

Participant ID	Geminate	n	Singleton	n	CC/C ratio
VenS1F	189 (51)	96	117 (41)	96	1.62
VenS2F	134 (25)	96	86 (26)	96	1.56
VenS3F	201 (33)	96	93 (23)	96	2.16
VenS1M	171 (28)	96	104 (29)	96	1.64
VenS2M	166 (26)	96	93 (26)	96	1.78
VenS3M	150 (27)	96	84 (26)	96	1.79
All	169 (40)	576	96 (31)	576	1.76

3.2. V1 duration

Table 4 shows the results for stressed V1. As expected, the geminate-singleton durational differences for pre-stress vowels were found to be non-significant. Therefore, only the results for the post-stress condition are reported.

Pre-geminate stressed V1 shortening occurred for all speakers, although to varying degrees. On average, the magnitude of the shortening was -28%, as revealed by the overall pre-geminate-to-pre-singleton vowel (Vcc/Vc) duration ratio reported in Table 4. Again, there is considerable variation across speakers.

The model found a significant overall effect of gemination on V1 duration across speakers (β = 51.14 ms, SE = 9.23, t = 13.83, p < .001). However, as mentioned above, post-hoc tests found significant geminate-singleton differences only for post-stress (β = -44.91 ms, SE = 3.67 ms, p < .001) and not for pre-stress (β = -8.72 ms, SE = 3.67 ms, p = 0.213) tokens.

Table 4. *Mean duration (in ms), SD (in parentheses), and counts (n) of stressed pre-geminate and pre-singleton V1 for each speaker. Vcc/Vc ratios are also provided.*

Post-stress condition (stressed V1)					
Participant ID	Geminate	n	Singleton	n	Vcc/Vc ratio
VenS1F	139 (22)	48	195 (35)	48	0.71
VenS2F	110 (14)	48	134 (19)	48	0.82
VenS3F	126 (18)	48	188 (21)	48	0.67
VenS1M	121 (19)	48	156 (15)	48	0.78
VenS2M	95 (13)	48	145 (21)	48	0.66
VenS3M	105 (17)	48	146 (22)	48	0.72
All	116 (23)	288	161 (32)	288	0.72

3.3. C/V1 ratio

Table 5 reports the results for C/V1 ratio. The effect of gemination on the model was significant across speakers and conditions (β = -0.77, SE = 0.23, t = -3.30, p < .05), although less highly so than for the previous two models based on actual durations.

Table 5. *Mean geminate and singleton C/V1 ratios, SD (in parentheses), and counts (n) for each speaker for the post-stress (top) and pre-stress (bottom) conditions.*

C/V1 ratios – Post-stress condition				
Part. ID	Geminate	n	Singleton	n
VenS1F	1.55 (0.56)	48	0.70 (0.38)	48
VenS2F	1.34 (0.33)	48	0.71 (0.27)	48
VenS3F	1.74 (0.50)	48	0.55 (0.14)	48
VenS1M	1.53 (0.51)	48	0.73 (0.24)	48
VenS2M	1.87 (0.50)	48	0.69 (0.32)	48
VenS3M	1.58 (1.17)	48	0.61 (0.23)	48
All	1.60 (0.67)	288	0.66 (0.27)	288

C/V1 ratios – Pre-stress condition				
Part. ID	Geminate	n	Singleton	n
VenS1F	2.64 (0.98)	48	1.53 (0.65)	48
VenS2F	2.29 (0.43)	48	1.31 (0.43)	48
VenS3F	3.09 (0.78)	48	1.11 (0.37)	48
VenS1M	2.87 (0.76)	48	1.50 (0.65)	48
VenS2M	3.39 (1.27)	48	1.62 (0.61)	48
VenS3M	2.41 (0.64)	48	1.33 (0.57)	48
All	2.78 (0.93)	288	1.40 (0.58)	288

At first glance the C/V1 ratio results seem to confirm those reported by [7] in that for post-stress tokens there appears to be a geminate-to-singleton cut-off ratio of 1.00 and for pre-stress tokens of 2.00. However, the SD values suggest that there might

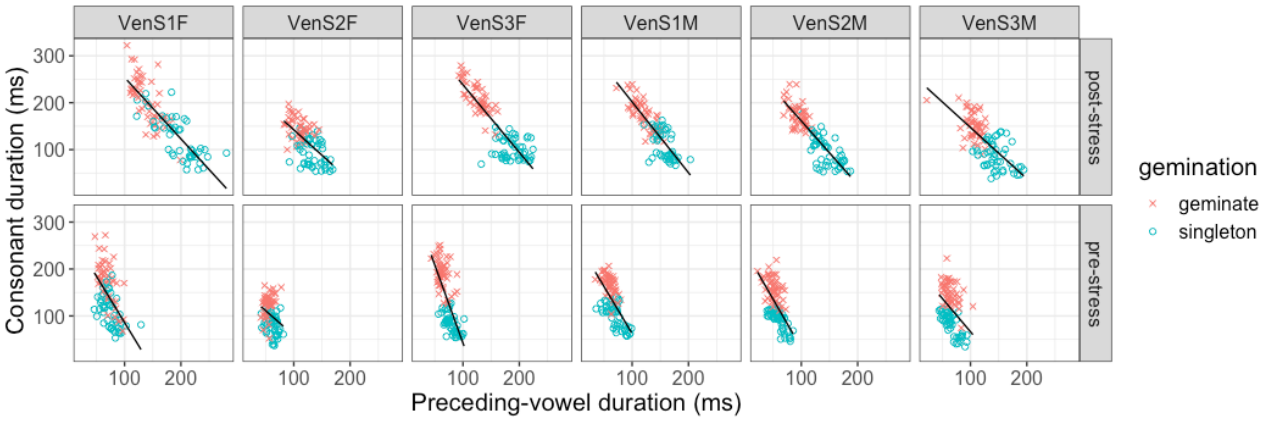


Figure 2: *C duration as a function of V1 duration by participant and stress condition. Lines of best fit are superimposed.*

be some overlap between categories for some speakers at around the assumed cut-off ratios. In fact, in contrast with the Roman variety [7], the relationship between C and V1 durations in this study appears to be linear and gradient for all our Veneto Italian speakers. Indeed, negative repeated-measures correlations [25] between the parameters were found for both stress conditions: very strong for the post-stress condition ($r = -0.814$, $p < .001$) and moderate-to-strong for the pre-stress condition ($r = -0.566$, $p < .001$). Figure 2 shows the particularly gradient nature of the C/V1 relationship found for this variety, with the data points roughly aligning around the line of best fit and the two data clusters appearing to merge into each other. This is in contrast with the strikingly more bimodal relationship observed in [7] for the Roman variety, where the data points form two clearly separated clusters with very little gradience and almost no overlap between length types. To give a better idea of the degree of overlap in this study, Table 6 shows how many of the singleton tokens would be categorized as geminate and vice-versa if the cut-off C/V1 ratios of 1.00 for post-stress and 2.00 for pre-stress tokens proposed by [7] were adopted.

Table 6. *Proportions of singleton tokens that would be categorized as geminate and vice-versa if the proposed cut-off C/V1 ratios were adopted.*

Participant ID	Sing. categorized as gem.	Gem. categorized as sing.
VenS1F	21/96 (21.9%)	26/96 (27.1%)
VenS2F	7/96 (7.3%)	16/96 (16.7%)
VenS3F	1/96 (1.0%)	6/96 (6.3%)
VenS1M	16/96 (16.7%)	13/96 (13.5%)
VenS2M	22/96 (22.9%)	4/96 (4.2%)
VenS3M	6/96 (6.3%)	17/96 (17.7%)
All	73/576 (12.7%)	82/576 (14.2%)

4. Discussion

In this study there is evidence of partial stop degemination for some Veneto Italian speakers. Partial degemination was measured in two ways: i) by comparing the observed CC/C ratios with those reported for the central varieties in similarly designed experimental settings; and ii) by looking at the amount of overlap in C/V1 ratios between length categories.

The above results also highlight noticeable speaker-specific differences which emerge more clearly when C/V1 is considered. Indeed, Figure 2 and Table 6 show a considerable degree of overlap between data points for most speakers except VenS3F. Recall that this speaker had been trained in Standard Italian pronunciation – something that becomes particularly evident here. By contrast, VenS1F exhibited particularly striking overlapping between length types. Recall that she was the oldest speaker and strongly favoured the use of the dialect. Furthermore, VenS1M – the second oldest participant – showed somewhat more overlap than the other two younger male speakers. It seems, therefore, that the sociophonetic factors listed in Table 1 may play a part in the phonetic realization of stops in Veneto speakers. However, only a purposely designed, large-scale sociophonetic experiment could provide more reliable information in this regard.

The differences between the findings of this study and those reported in [7] for the Roman variety suggest that the phonetic realization of gemination may vary between the two varieties more in terms of the relative contributions of C and V1 duration than in terms of absolute C duration. Whereas in the Roman variety geminate consonants are consistently longer and pre-geminate vowels consistently shorter, in this study the contrast between the two length categories appears to be more blurred. The consonants exhibited greater durational variability and the duration of the preceding vowels co-varied accordingly in a particularly gradient fashion. This resulted in a distinctly gradient inverse relationship between C and V1 durations which does not allow for a reliable clear-cut distinction between geminates and singletons based on C/V1 ratios, as is the case with the Roman variety and presumably other Central-Southern varieties. Furthermore, it is possible that the blurriness between length categories due to gradient C/V1 ratio distribution may contribute to listeners' perception of degemination in Veneto Italian, although this would have to be investigated in a future perceptual experiment.

5. Conclusion

This study suggests that partial degemination may be the norm for adult Veneto Italian speakers, with speaker-specific differences possibly linked to sociophonetic factors such as speaker age, educational background, and level of Italian-dialect bilectalism. It appears that older, less educated Veneto Italian speakers who more strongly favour the use of the dialect may be particularly prone to exhibiting relatively higher degrees of degemination.

6. References

[1] Bertinetto, P.M., and Loporcaro, M., "The sound pattern of Standard Italian, as compared with the varieties spoken in Florence, Milan and Rome", *JIPA*, vol. 35, no. 2, pp. 131–151, Dec. 2005, doi: 10.1017/S0025100305002148.

[2] Rogers, D., and D'Arcangeli, L., "Italian", *JIPA*, vol. 34, no. 1, pp. 117–121, Jan. 2004, doi: 10.1017/S0025100304001628.

[3] Payne, E. M., "Phonetic variation in Italian consonant gemination", *JIPA*, vol. 35, no. 2, pp. 153–181, Dec. 2005, doi: 10.1017/S0025100305002240.

[4] Fava, E., and Magno Caldognetto, E., "Studio sperimentale delle caratteristiche elettroacustiche delle vocali toniche ed atone in bisillabi italiani", in *Studi di fonetica e fonologia*, pp. 35–79, Società di Linguistica Italiana, 1976.

[5] Rossetti, R., "Gemination of Italian stops", *JASA*, vol. 95, no. 5, pp. 2874–2874, May 1994, doi: 10.1121/1.409450.

[6] Esposito, A., and Di Benedetto, M. G., "Acoustical and perceptual study of gemination in Italian stops", *JASA*, vol. 106, no. 4, pp. 2051–2062, Oct. 1999, doi: 10.1121/1.428056.

[7] Pickett, E. R., Blumstein, S. E., and Burton, M. W., "Effects of speaking rate on the singleton/geminate consonant contrast in Italian", *Phonetica*, vol. 56, no. 3–4, pp. 135–157, Dec. 1999, doi: 10.1159/000028448.

[8] Cerrato, L., and Falcone, M., "Acoustic and perceptual characteristics of Italian stop consonants", in *ICSLP'98: Proceedings of the 5th International Conference on Spoken Language Processing*, Sydney, 1998.

[9] Hualde, J. I., and Nadeu, M., "Lenition and phonemic overlap in Rome Italian", *Phonetica*, vol. 68, no. 4, pp. 215–242, Jan. 2012, doi: 10.1159/000334303.

[10] Di Benedetto, M. G., Shattuck-Hufnagel, S., De Nardis, L., Budoni, S., Arango, J., Chan, I., DeCaprio, A., "Lexical and syntactic gemination in Italian consonants—Does a geminate Italian consonant consist of a repeated or a strengthened consonant?", *JASA*, vol. 149, no. 5, pp. 3375–3386, May 2021, doi: 10.1121/10.0004987.

[11] Chang, W., "Geminate vs. non-geminate consonants in Italian: Evidence from a phonetic analysis", *Uni. Penn. WPL*, vol. 7, no. 1, pp. 53–63, 2000.

[12] Farnetani, E., and Kori, S., "Effects of syllable and word structure on segmental durations in spoken Italian", *Speech Communication*, no. 5, pp. 17–34, 1986.

[13] Hajek, J., Stevens, M., and Webster, G., "Vowel duration, compression and lengthening in stressed syllables in Italian", in *Proceedings of the 16th International Congress of Phonetics Sciences*, Saarbrücken, vol. ICPhS XVI, pp. 1057–1060, 2007.

[14] Hajek, J., and Stevens, M., "Vowel duration, compression and lengthening in stressed syllables in Central and Southern varieties of Standard Italian", *Interspeech 2008*, vol. 22, pp. 516–519, 2008.

[15] Celata, C., Meluzzi, C., and Bertini, C., "Acoustic and kinematic correlates of heterosyllabicity in different phonological contexts", *Language and Speech*, pp. 1–26, Jan. 2022, doi: 10.1177/00238309211065789.

[16] Canepari, L., *Lingua italiana nel Veneto*. Padova: CLESP, 1984.

[17] Canepari, L., and Giovannelli, B., *La buona pronuncia italiana del terzo millennio: manualetto d'italiano neutro con sonori, esercizi e test*, 4. edition. Rome: Aracne, 2012.

[18] Canepari, L., *Manuale di pronuncia italiana: con un pronunciario di oltre 30,000 voci e due audiocassette C45*, 1. ed. Bologna: Zanichelli, 1992.

[19] Mairano, P., and De Iacovo, V., "Gemination in Northern versus Central and Southern varieties of Italian: A corpus-based investigation", *Language and Speech*, vol. 63, no. 3, pp. 608–634, Sep. 2020, doi: 10.1177/0023830919875481.

[20] Sanfelici, E., and Roch, M., "The native speaker in Italian-dialects bilingualism: Insights from the acquisition of Vicentino by preschool children", *Frontiers in Psychology*, vol. 12, art. 717639, Oct. 2021, doi: 10.3389/fpsyg.2021.717639.

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

[22] Winkelmann, R., Jaensch, K., Cassidy, S., and Harrington, J., *emuR: Main package of the EMU Speech Database Management System*. 2021.

[23] Kuznetsova, A., Brockhoff, P. B., and Christensen, R. H. B., "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.

[24] Lenth, R.V., Buerkner, P., Herve, M., Jung, M., Love, J., Miguez, F., Riebl, H., Singmann, H., "Package 'emmeans'", 2022. [Online]. Available: https://cran.r-project.org/web/packages/emmeans/emmeans.pdf.

[25] Bakdash, J.Z., and Marusich, L.R. "Repeated Measures Correlation", *Frontiers in Psychology*, vol. 8, art. 456, 2017, doi: 10.3389/fpsyg.2017.00456.

Assessing the validity of remote recordings captured with a generic smartphone application designed for speech research

Joshua Penney, Ben Davies, Felicity Cox

Centre for Language Sciences, Department of Linguistics, Macquarie University
joshua.penney@mq.edu.au; ben.davies@mq.edu.au; felicity.cox@mq.edu.au

Abstract

This paper introduces a generic smartphone application designed for collecting speech data remotely: Appen Research. Data collected from 24 female Australian English-speaking participants using the smartphone application were compared to data collected from the same participants with laboratory-based equipment. F1 and F2 values were extracted from the midpoints of the 11 stressed monophthongs of Australian English. While some non-low vowels showed slight raising of F1 values in the data recorded with the application, overall the results suggest that recordings collected with the smartphone application are generally comparable to recordings made in a studio for the purposes of analysing vowel formants.

Index Terms: remote data collection, smartphones, vowels, Australian English, monophthongs, recording methods

1. Introduction

1.1. Background

There has been a shift in recent years towards the use of simple personal electronic devices such as smartphones, tablets, and laptop computers to facilitate data collection for scientific research as such portable devices have become increasingly common (at least within certain sections of the population) [1, 2, 3, 4]. With regard to speech research, an approach to data collection that leverages the accessibility of such devices which typically contain an inbuilt microphone and an internet connection makes economical and practical sense. For example, in Europe a number of custom-made recording applications (apps) for smartphones that were paired with targeted (social) media campaigns have successfully exploited high levels of mobile phone ownership to crowdsource speech data from large numbers of speakers of certain languages and dialect groups [5, 6, 7, 8, 9, 10].

During the COVID-19 pandemic, public health orders resulted in restrictions on face-to-face contact as well as restraints on movement, at times with various state and national borders being closed. Under these circumstances, traditional laboratory-based data collection was generally unable to proceed, and as a result many researchers turned to remote data collection in order to continue their research. This resulted in a surge of interest in options for remote data collection, as well as a number of comparison studies examining the viability of data collected remotely for speech research and the extent to which acoustic measurements are affected by such methods [11, 12, 13, 14, 15]. These studies complement research assessing the validity of speech data collected via smartphones for clinical voice analysis [16, 17, 18]. Although recordings made with personal devices tend to show some deviation from those made with reference devices, in general most studies suggest that recordings made with modern smartphones are sufficiently similar to those made with traditional lab-based recording equipment, at least for the purposes of examination of F0 and the first two-three formants [11, 14, 18]. There are some important caveats to ensure that the quality of the recordings is sufficient for acoustic data analysis: data should be captured in lossless rather than lossy formats (e.g. wav rather than mp3), should not be recorded over an internet connection but rather locally recorded (i.e. on a participant's phone) prior to data being transferred, and deviations may be more problematic in particular vowels and in particular individuals [11, 13, 14]. Furthermore, for some fine-grained analyses, remote recordings made with smartphones may not be suitable [11, 15]; for example, [15, 18] found problematic differences for some acoustic measurements of voice quality. This was particularly the case for voice quality measurements based on harmonic amplitudes.

In the Australian context, remote data collection offers great potential to speech researchers. It may lead to increased participation, as it removes barriers such as the need to travel to and be present physically at a university lab; rather, participation can be done in the comfort of a participant's own home at a time that is convenient for them [12, 19]. This has the potential to open up participation to a wider sample of the population, rather than the overrepresentation of (mostly young, middle class, female) university students that very often form the participant base in many speech studies.

The ability to record participants remotely is particularly important given the geography of Australia, as it could enable speakers from rural and isolated communities outside of major urban centres to be better represented in speech research. For example, this would be beneficial for work examining variation within languages, and would allow comparisons to be made between speakers from more than just a subset of locations [20, 21, 22]. Similarly, the use of remote data collection methods could assist in documentation and revitalisation of under-studied languages such as Australian languages spoken in remote indigenous communities, with recording apps having been shown to be a useful tool for research on endangered languages [23, 24]. Remote data collection would not only provide benefits for researchers working with populations located within Australia; researchers who work on and with speech communities in overseas countries could also benefit, as the ability to record participants remotely could reduce the number of fieldwork trips and the associated travel costs that are generally required for in person data collection [12], and could also ease issues relating to accessibilty, which may be a challenge faced by some researchers. As alluded to above, the ability to collect speech data remotely can also serve to ensure that research can be continued in the case of renewed restrictions on movement, whether due to public health restrictions or other unforeseen events.

As mentioned above, some speech research teams have created their own custom recording apps that are specially designed to address particular research questions by eliciting speech through highly controlled tasks [5, 6, 7, 8, 9, 10, 23]. While such an approach is ideal for addressing specific research questions, designing and programming a custom made app is also costly, time-consuming, and requires advanced programming abilities (or funding to employ external programmers). This puts such an approach beyond the reach of many researchers, particularly students and early career researchers. Additionally, as smartphone operating systems are continually updated, ongoing maintenance is required to ensure compatibility with new operating system versions, without which a custom app may become deprecated over a short time period. Moreover, as such apps are designed with specific languages/dialects/research questions in mind, the result is that multiple researchers create a variation of what is, essentially, the same tool over and over again at great effort and expense.

The issues outlined above highlight the need for a generic recording app to enable remote speech data collection across many different projects. Some generic recording apps do exist; however, there is a distinct lack of simple, cross-platform tools that are able to capture the high quality, uncompressed recordings necessary for speech research. Moreover, existing apps that provide high quality recordings generally require subscriptions or come with advertisements. In most cases participants need to manage data settings (e.g. file formats, sampling rates, etc.) and then manually upload data to a repository, or in some cases email the saved files to the researcher: a dangerously insecure data management practice that inevitably leads to suboptimal data quality and missing – or in a worst case scenario, stolen – data.

In this paper, we introduce the Appen Research app, a simple, easy to use, generic recording app that has been designed specifically with the requirements of speech researchers in mind.

1.2. Appen Research

The Appen Research app has been designed to function as a simple, cross-platform smartphone recording device that can be used for remote data collection of high-quality speech data. The app is currently in beta testing mode. In its current configuration, it can record speech to wav file format with a 44.1kHz sampling rate and 16-Bit resolution with a maximum file duration of 15 minutes. In addition to the recorded files, metadata related to smartphone make and model, and operating system and version are also collected with each recording.

Importantly, the app is not linked to a particular experiment or project, nor does it include specific instructions for a participant to follow or consent/questionnaire forms to complete. This enables it to be used for a variety of data collection purposes on any number of different projects. For example, the app could be used to capture data from participants in tasks that are supervised by a researcher, in which supervision could be carried out by video/phone call (on a separate device). In another case, a participant may be given instructions for a simple reading task that should be completed at regular intervals without researcher supervision. The app could be used for highly controlled experiments, for recording conversations between two interlocutors, or for capturing infant direct speech. In some cases, both local and non-local participants may take part in the same study, and recording via the app would ensure a consistent recording methodology between participants. It is entirely up to the researcher how to implement their task; the app is envisaged merely as a 'portable' recording device.

Prior to data collection beginning, the researcher needs to establish a project identification code and a set of participant identification numbers to allocate to participants. Additionally, the researcher needs to designate a secure storage location for the data to be transferred to. Currently, data is configured to be transferred directly to a user's CloudStor account using the FileSender API. The recorded file is encrypted on the participant's device, temporarily staged on Appen's servers and uploaded to CloudStor within a 15 minute window. The use of end-to-end encryption is vital when collecting data remotely. It gives researchers control over their data and ensures participants can trust their recordings are not accidentally released to the wider public or stolen by those with malicious intent. AARNet's CloudStor/FileSender environment was selected to perform this function within the Appen Research app because the company is owned by a consortium of Australian universities, making it highly sensitive to the evolving data and privacy concerns of Australian research institutes.

The app is designed for simple participant usability. It has cross-platform compatibility with both Apple iOS and Android operating systems, and will be available for free for participants to download onto their own smartphone. Once a participant enters a project identification code and their participant identification number, no settings need to be adjusted; rather, the participant's interface consists of a record/pause/stop button to begin, pause, and end the recording. Once the recording is completed, the participant is prompted to upload and transfer the file, after which the task is comlete. Data transfer only takes place when a wifi connection is available.

In this paper, we present the preliminary results of a study designed to compare vowel realisations of the Australian English (AusE) monophthongs in recordings captured with the Appen Research app against recordings of the same population recorded in a laboratory setting with professional recording equipment.

2. Methods

2.1. Participants

We recruited 25 female speakers (aged between 19 and 64; mean age: 28) to take part in this study. Of these, 20 were undergraduate students in the Department of Linguistics at Macquarie University, who received course credit for their participation. The remaining five participants were researchers in the Department of Linguistics at Macquarie University, who were not compensated for their time. All participants were L1 or early L2 speakers of AusE who had completed all of their schooling in Australia. Data for one participant were excluded due to a technical issue that resulted in a partial loss of data, leaving data for 24 speakers remaining for analysis.

2.2. Procedure

All participants completed the same task in two successive sessions recorded in a sound-treated room in the Department of Linguistics at Macquarie University: in one of the sessions, participants were recorded with a Neumann TL103 condenser microphone using open-source recording software Audacity (https://www.audacityteam.org/) with a sampling rate of 44.1 kHz and 16-Bit resolution. Recordings from this session will hereafter be referred to as *Studio* data. In the other session, participants were recorded through the Appen Research app

with a sampling rate of 44.1 kHz and 16-Bit resolution, which they were instructed to install on their own personal smartphone prior to the session. Recordings from this session will hereafter be referred to as *App* data.

Participants were recorded as they read aloud 90 sentences presented orthographically on a computer monitor. Each sentence contained a monosyllabic target word with the standard /hVd/ structure, where /V/ comprised the 18 stressed vowels of AusE [23]. The target words were embedded within a carrier sentence with the form: *say <TARGET> again*, with speakers instructed to read the sentences casually as if speaking to a friend. For each participant, five repetitions of each of the 18 stressed vowels were sampled. The order of presentation was randomised for each participant (with the same presentation order in both sessions). It was our intention to counterbalance the order of sessions; however, as data collection is ongoing and the analysis here is based on an initial subset of collected data, the data examined here are not equally balanced (*Studio* prior to *App*: 20; *App* prior to *Studio*: 5). For the purposes of this analysis, only data relating to the 11 AusE monophthongs were examined.

2.3. Acoustic analysis

The collected files were orthographically transcribed and subsequently automatically segmented and force-aligned through WebMAuS [26] utilising an AusE model. The resulting textgrids and corresponding sound files were converted to an emu database using emuR [27].

Formant frequencies were calculated with Praat [28] and imported into the database via PraatR [29]. Formant frequencies were calculated for all back and central vowels with the default settings: Max. formant: 5500Hz; No. formants: 5, Window length: 0.025s; front vowels were estimated with the following settings as these resulted in improved formant tracking (and consequently fewer outliers): Max. formant: 6600Hz; No. formants: 5, Window length: 0.025s.

All files were inspected and segment boundaries for the vowels were hand corrected where necessary. In some cases, participants misread the sentence and produced an incorrect target word: 37 files were excluded from analysis due to such errors. F1 and F2 measurements (in Hertz) were then extracted at the temporal midpoint of each vowel. Outliers of each vowel were subsequently trimmed using the modified Mahalanobis distance method [30]. This resulted in 2472 files remaining for the analysis (*App*: 1243; *Studio*: 1229).

2.4. Statistical analysis

Potential differences between the *Studio* and *App* files were examined using linear mixed effects regression models with the lme4 [31] package in R [32], with *p*-values calculated by likelihood ratio tests with the afex [33] package. Separate models were fitted for F1 and F2, in each case with the formant in question included as the dependent variable. Fixed factors were Recording method (i.e. *Studio* vs *App*) and Vowel. We also included an interaction term between these fixed factors. Random intercepts were included for Participant. Random slopes were included for Recording method by Participant in the F1 model. The inclusion of Recording method by Participant in the F2 model and the inclusion of random slopes for Vowel by Participant in both F1 and F2 models resulted in singular model fits; hence these were not included in the final models. Post-hoc pairwise comparisons were conducted with the emmeans package [34] with Tukey HSD corrections for multiple comparisons.

As participants utilised their own personal smartphone for the *App* recordings, it was not possible for us to control for or balance the operating system and version of the operating system that was used. The majority of participants (n = 19) completed the *App* session using an Apple iPhone and a version of the Apple iOS operating system. The remaining participants used a Samsung (n = 4) or Google (n = 1) smartphone and a version of the Android operating system. Despite operating system not being well sampled, we included this as a covariate in our initial modelling. However, this was not a significant effect for either F1 or F2 and did not improve the model fit and was therefore removed from the final models. In addition, a subsequent analysis of the data excluding participants who used the Android operating system did not show substantial differences to the analysis that included all participants.

3. Results

Figure 1 illustrates the distribution of vowels according to the two recording methods, as presented in the traditional F1/F2 vowel plane. Means of each of the vowels are shown by the vowel labels, and the ellipses represent 95% confidence intervals. As can be seen, the monophthongal space appears similar in the data captured by the two recording methods and the vowels appear to be comparably dispersed. However, some small differences can also be noticed; in particular, some of the vowels in the *Studio* data – particularly the non-low vowels – appear to be slightly raised, i.e. they appear to be lower in F1.

The linear mixed effects model for F1 showed significant effects for Recording method ($F(1, 22) = 8.11$; $p = 0.009$) and Vowel ($F(10, 2405) = 3456.67$; $p < 0.001$), as well as their interaction ($F(10, 2406) = 4.66$; $p < 0.001$). Post-hoc pairwise comparisons revealed that the only vowel that differed significantly between the recording methods was /iː/ ($p = 0.041$), which had a mean F1 value in the *Studio* data that was 28.4 Hz lower than in the *App* data. Pairwise comparisons further showed that within each recording method each vowel differed significantly in F1 from all other vowels ($p = 0.0079$ or below) with the exception of /iː/ and /ʉː/ and /ɪ/ and /ʊ/ in the *App* data, and /iː/ and /ʉː/, /ɪ/ and /ʊ/, and /e/ and /ɔ/ in the *Studio* data. That is, /e/ and /ɔ/ differed significantly from one another in the App data (with a difference of 29.4 Hz), but not in the Studio data (with a difference of 20.7 Hz), where mean values for /ɔ/ were marginally higher.

The linear mixed effects model for F2 showed a significant effect for Vowel ($F(10, 2427) = 6678.36$; $p < 0.001$). There was no effect of Recording method or the interaction between Recording method and Vowel for F2. Post-hoc pairwise comparisons showed that each of the vowels differed significantly in F2 from all other vowels (all $p = 0.0013$ or below), apart from /ʊ/ and /ɔ/.

4. Discussion

Overall, the results above show that differences in formant measurements between the two recording methods are relatively minimal, which is consistent with previous findings [14, 18]. Small differences in F1 were observable for some vowels, with lower mean F1 values (corresponding to higher vowel realisations) in the *Studio* data than in the *App* data. However, this effect was only found to be significant for /iː/, and even in that case the size of the effect was not considerable. Given the fact that /iː/ displays variable onglide in AusE [20], it may be possible that the greater difference in this vowel could be due to measurements being taken at the vowel midpoint

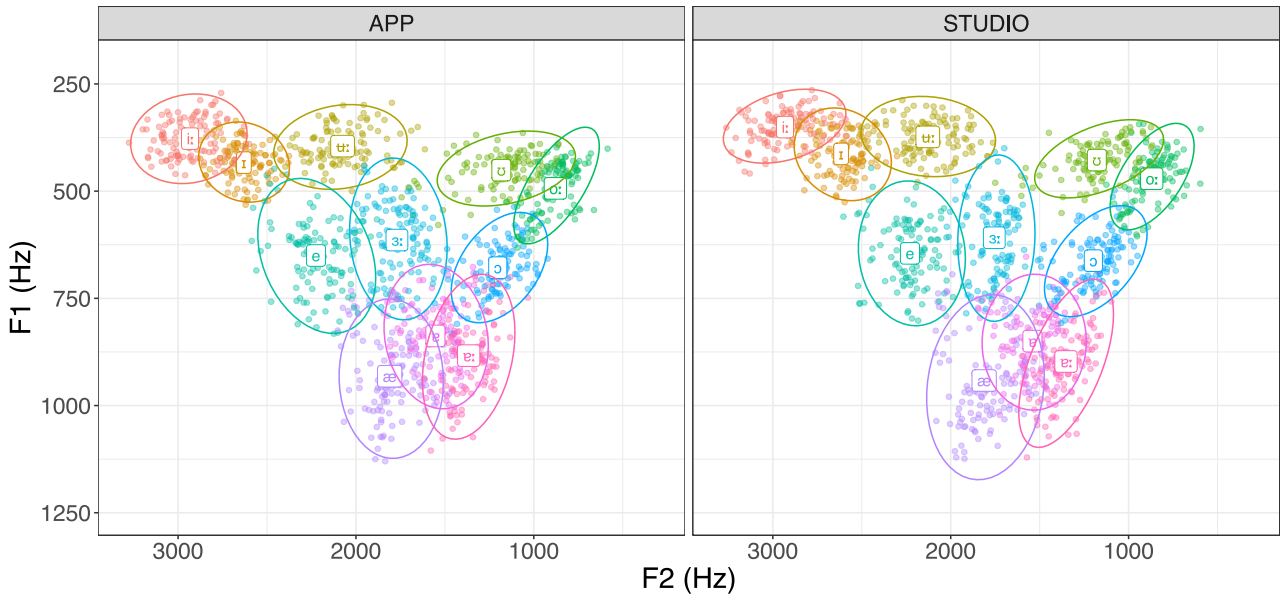


Figure 1. F1 and F2 values of all monophthongs in the *App* (left panel) and *Studio* (right panel) data. Text labels represent mean values. Ellipses represent 95% confidence intervals.

rather than identifying the vowel target. In both recording methods, vowels were found to be similarly dispersed, with the only difference found relating to /e/ and /ɔ/, which showed a significant difference in F1 in the *App* data, but not in the *Studio* data, where there was a higher realisation (lower mean F1 value) of /ɔ./ Taken together, these results suggest that F1 values in the recordings made with the Appen Research app may be affected by a slight raising; that is, some vowels may be measured as marginally lower when captured with the app.

This apparent effect of F1 raising on some (non-low) vowels in the *App* recordings may be the result of internal digital signal processing within the smartphones used by the participants. Most smartphones employ a range of algorithms to digitise and improve the audio signal, for example to reduce background noise and enhance the clarity of the speech that is captured [35, 36]. These processes, and the extent of processing, vary between different makes and models of smartphone, with little specific information available publicly about exactly what processes are applied in a particular phone. Nevertheless, such processes may very well impact upon acoustic measurements to some extent [37]. Band-pass filters that have traditionally been used in the transmission of speech by telephone are also known to inflate F1 values in non-low vowels in some cases [38, 39], among other effects. However, the effects seen here do not appear to be as substantial as those previously reported, and appear to be limited to F1: there were no significant differences according to F2.

Note that it was not our intention here to compare recordings of the same utterances captured simultaneously with different recording devices, as is often the case in studies comparing speech recorded with personal devices. Rather, in this study it was our intention to assess whether recordings of the same general population would yield effectively comparable results in non contemporaneous recordings. Apart from the small differences discussed above, this appears to be the case. We therefore suggest that remote data collection with the Appen Research app may be an effective means for conducting speech research in the Australian context, at least for the examination of F1 and F2 values in AusE monophthongs. Of course, recordings captured with the app may be more suitable for some lines of enquiry rather than others, as some acoustic measurements are likely to be more susceptible to deviations in recordings made with smartphones [11, 13, 15].

It should be pointed out that this preliminary study was based on a relatively small sample size. As noted above, utilising smartphones for remote recordings may facilitate greater participation, which in turn would lead to larger sets of data for analysis. Further research will determine whether the findings shown here also hold over a greater number of participants and with speakers from more heterogeneous backgrounds.

5. Conclusion

This paper introduced the Appen Research app, a generic smartphone based recording application for speech research, and has shown that recordings from participants' smartphones captured with this smartphone application are generally comparable to recordings made in a studio for the purposes of analysis of the first two formants taken from the midpoints of AusE monophthongs. The use of this application may therefore be of benefit to researchers interested in collecting data remotely with the intention of examining measurements of vowel formants. Future work will address the extent to which other acoustic measurements are also comparable between remote and studio-based recordings.

6. Acknowledgements

We thank members of the MQ phonetics lab and participants in the Challenges for Change workshop at LabPhon18 for their comments and suggestions. This work was supported by ARC Future Fellowship Grant FT180100462 to the third author.

7. References

[1] Birenboim, A. and Shoval, N., "Mobility research in the age of the smartphone," Ann. Am. Assoc. Geogr., 106(2):283–291, 2016.

[2] Harari, G. M., Lane, N. D., Wang, R., Crosier, B. S., Campbell, A. T., and Gosling, S. D., "Using smartphones to collect behavioral data in psychological science: opportunities, practical considerations, and challenges," Perspect. Psychol. Sci. 11, 838–854, 2016.

[3] Miller, G., "The Smartphone Psychology Manifesto," Perspect. Psychol. Sci. 7(3): 221–237, 2012.

[4] Seifert Alexander, Hofer Matthias, Allemand Mathias., "Mobile data collection: Smart, but not (yet) smart enough," Frontiers in Neuroscience, 12, 2018. DOI=10.3389/fnins.2018.00971

[5] Leemann, A., Kolly, M.-J., Goldman, J.-P., Dellwo, V., Hove, I., Almajai, I., and Wanitsch, D., "Voice Äpp: A mobile app for crowdsourcing Swiss German dialect data," in Proc. INTERSPEECH 2015, Dresden, 2804–2808, 2015.

[6] Leemann, A., Kolly, M.-J., and Britain, D., "The English Dialects app: The creation of a crowdsourced dialect corpus," Ampersand, 5: 1–17, 2018.

[7] Entringer, N., Gilles, P., Martin, S., and Purschke, C., "Schnëssen: Surveying language dynamics in Luxembourgish with a mobile research app," Linguistics Vanguard, 7(s1): 1–15, 2021.

[8] Gittelson, B., Leemann, A., and Tomaschek, F., "Using crowd-sourced speech data to study socially constrained variation in nonmodal phonation," Frontiers in Artificial Intelligence, 3: 1–9, 2021.

[9] Hilton, N. H., "Stimmen: A citizen science approach to minority language sociolinguistics," Linguistics Vanguard, 7(s1):1–15, 2021.

[10] Leemann, A., "Apps for capturing language variation and change in German-speaking Europe: Opportunities, challenges, findings, and future directions," Linguistics Vanguard, 7(s1): 1–12, 2021.

[11] Freeman, V., P. DeDecker, and M. Landers, "Suitability of self recordings and video calls: Vowel formants and nasal spectra," J. Acoust. Soc. Am, 148: 2714, 2020.

[12] Leemann, A., Jeszenszky, P., Steiner, C., Studerus, M. and Messerli, J., "Linguistic fieldwork in a pandemic: Supervised data collection combining smartphone recordings and video conferencing," Linguistics Vanguard, 6(s3): 1–16, 2020

[13] Sanker, C., Babinski, S., Burns, R., Evans, M., Johns, J., Kim, J., Smith, S., Weber, N., Bowern, C., "(Don't) try this at home! The effects of recording devices and software on phonetic analysis," Language, 97(4): e360-e382, 2021.

[14] Zhang, C., Jepson, K., Lohfink, G., and Arvaniti, A., "Comparing acoustic analyses of speech data collected remotely," J. Acoust. Soc. Am., 149: 3910–3916, 2021.

[15] Penney, J., Gibson, A., Cox, F., Proctor, M., & Szakay, A., "A comparison of acoustic correlates of voice quality across different recording devices: A cautionary tale," in Proc. INTERSPEECH 2021, Brno, 1389–1393, 2021.

[16] Manfredi, C., Lebacq, J., Cantarella, G., Schoentgen, J., Orlandi, S., Bandini, A., and DeJonckere, P. H., "Smartphones offer new opportunities in clinical voice research," J. Voice, 31(1): 111.e1–111.e7, 2016.

[17] Grillo, E. U., Brosious, J. N., Sorrell, S. L., and Anand, S.,"Influence of smartphones and software on acoustic voice measures," Int. J. Telerehabilitation, 8: 9–14, 2016.

[18] Jannetts, S., Schaeffler, F., Beck, J., and Cowen, S., "Assessing voice health using smartphones: Bias and random error of acoustic voice parameters captured by different smartphone types," Int. J. Lang. Commun. Disord., 54(2): 292–305, 2019

[19] Archibald, M. M., Ambagtsheer, R. C., Casey, M. G., and Lawless, M., "Using Zoom video conferencing for qualitative data collection: Perceptions and experiences of researchers and participants," Int. J. Qual. Methods, 18: 1–8, 2019.

[20] Cox, F., and Palethorpe, S., "The border effect: Vowel differences across the NSW/Victorian border," In C. Moskovsky [Ed.], *Proc. Aust. Ling. Soc 2003,* 1–14, 2004.

[21] Burnham, D., Estival, D., Fazio, S., Viethen, J., Cox, F., Dale, R., Cassidy, S., Epps, J., Togneri, R., Wagner, M., Kinoshita, Y., Göcke, R., Arciuli, J., Onslow, M., Lewis, T., Butcher A., and Hajek, J., "Building an audio-visual corpus of Australian English: Large corpus collection with an economical portable and replicable black box," in Proc. INTERSPEECH 2011, Florence, 841–844, 2011.

[22] Cox, F., and Palethorpe, S., "Vowel variation across four major Australian cities," in Proc. ICPhS, Melbourne, 577–581, 2019.

[23] Bird, S., Hanke, F. R., Adams, O., and Lee, H., "Aikuma: A mobile app for collaborative language documentation," Proc. Workshop on the Use of Computational Methods in the Study of Endangered Languages, 1–5, 2014.

[24] Bird, S., "Designing Mobile Applications for Endangered Languages," in K. L. Rehg and L. Campbell [Eds], The Oxford Handbook of Endangered Languages, Oxford University Press, 2018.

[25] Cox, F., and Palethorpe, S., "Australian English", J. Int. Phon. Assoc., 37(3): 431–350, 2007.

[26] Kisler, T., Reichel, U., and Schiel, F., "Multilingual processing of speech via web services," Comput. Speech Lang., 45: 326–347, 2017.

[27] Winkelmann, R., Harrington, J., and Jänsch, K., "EMU-SDMS: Advanced speech database management and analysis in R," Comput. Speech Lang., 45: 392–410, 2017.

[28] Boersma, P., and Weenink, D., "Praat: Doing phonetics by computer," version 6.1.16, 2020 [Computer program]. Available: http://www.praat.org/

[29] Albin, A. L. "PraatR: An architecture for controlling the phonetics software "Praat" with the R programming language," J. Acoust. Soc. Am., 135(4): 2198, 2014.

[30] Stanley, J. A. "The Absence of a Religiolect among Latter-Day Saints in Southwest Washington," In V. Fridland, A. Beckford Wassink, L. Hall-Lew, and T. Kendall [Eds], Speech in the Western States: Vol. 3, Understudied Varieties, 95–122. Duke University Press, 2020.

[31] Bates, D., Mächler, M., Bolker, B., and Walker, S., "Fitting linear mixed-effects models using lme4," J. Stat. Softw., 67(1), 2015.

[32] R Core Team, "R: A language and environment for statistical computing," version 4.0.2, 2020 [Computer program]. Available: https://www.r-project.org/

[33] Singmann, H., Bolker, B., Westfall, J., Aust F., and Ben-Shachar, M. S., "afex: Analysis of Factorial Experiments," version 1.0-1, 2021 [R package]. Available: https://CRAN.R-project.org/package=afex

[34] 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

[35] Tan, K., Zhang, X., and Wang, D., "Real-time speech enhancement using an efficient convolutional recurrent network for dual-microphone mobile phones in close-talk scenarios," in Proc. ICASSP 2019, 5751–5755, 2019.

[36] Teutsch, H., "Audio and Acoustic Signal Processing's Major Impact on Smartphones," IEEE Signal Processing Society, Online: https://signalprocessingsociety.org/publications-resources/blog/audio-and-acoustic-signal-processing%E2%80%99s-major-impact-smartphones, Accessed 20 June, 2022.

[37] Faber, B. M., "Acoustical measurements with smartphones: Possibilities and limitations," Acoustics Today, 13(2): 10–17, 2017.

[38] Künzel, H.J., "Beware of the 'telephone effect': The influence of telephone transmission on the measurement of formant frequencies," Int. J. Speech Lang. Law, 8(1): 80–99, 2001.

[39] Byrne, C., and Foulkes, P., "The 'mobile phone effect' on vowel formants," Int. J. Speech Lang. Law, 11(1): 83–102, 2004.

Young Aucklanders and New Zealand English Vowel Shifts

Brooke Ross, Elaine Ballard and Catherine Watson

The University of Auckland

bros138@aucklanduni.ac.nz, e.ballard@auckland.ac.nz, c.watson@auckland.ac.nz

Abstract

Two distinctive features of New Zealand English are the short front vowel shift, and the articulation of the NURSE vowel. These shifts have resulted in raised pronunciations of DRESS, TRAP, NURSE. Until recently, New Zealand English research has suggested that these changes are ongoing with little regional variation. In the following paper we report findings from a study with Auckland based New Zealand English speakers which suggest this might be changing. Sociolinguistic interviews were recorded with 67 Auckland based participants. Speakers are stratified by age (16-25 and 40+) and gender. Hand-corrected formants from stressed vowels, marked at the vowel target, were analyzed. The analysis looks at over 20,000 monophthong tokens. The results suggest there are differences between the older and younger speakers. The younger speakers have lowered and retracted DRESS, TRAP and NURSE vowels. Although some results are confounded by the effects of aging on the vocal tract. The implication of these results is discussed.

Index Terms: New Zealand English, Auckland, Vowels, Acoustic Analysis, Sound Change, Monophthongs

1. Introduction

Two distinctive features of New Zealand English (NZE) are the short front vowel shift and the articulation of the NURSE vowel. The first results in raised DRESS and TRAP vowels, and a lowered and retracted KIT vowel [1]. In addition, the raising of DRESS has also resulted in the diphthongization of the long FLEECE vowel [2,3]. The second is characterized by raising of the NURSE vowel towards GOOSE [4]. Until recently, New Zealand English phonetic research has concluded that these shifts are ongoing in NZE, with continued DRESS raising and a greater onglide for FLEECE [1, 2, 3, 5, 6]. These claims have been made, however, assuming that New Zealand English lacks regional variation [7,8].

While this may be historically true, this may no longer be the case, with the most recent research on NZE in New Zealand's largest city Auckland [9, 10]. Auckland has undergone notable demographic change with over 40% of its residents born overseas [11]. For some of these migrants English is a new language, while for others a different variety of English is the spoken norm. As noted in Cheshire and other European research, such linguistic diversity can be a catalyst for linguistic change [12].

The phonetic analysis of young NZE speakers in Auckland [9, 10], contrary to previous NZE research, found TRAP and DRESS lowering and reduced FLEECE diphthongization. As the findings from this study were limited to read speech from young speakers, the conclusions reached in this analysis were preliminary in nature. The current study expands on that research by acoustically analyzing the speech of older Aucklanders together with younger speakers and by focusing on conversations rather than read speech from these speakers. In addition to analyzing changes in the short vowels the study will also consider the NURSE vowel. Widening the lens to look beyond the short front vowels, and to investigating the natural speech of both young and old speakers in Auckland, will provide us with more reliable apparent time evidence of sound change in Auckland.

2. Method

2.1. Speakers

This analysis uses data collected for the Auckland Voices Project (details in [9]). It consists of 67 NZE speakers from Auckland, stratified by age and gender (Older n=29 (17 women, 12 men), Younger n=39 (20 women, 18 Men)). Participants were recorded in a sociolinguistic style interview for 1-2 hours in a quiet location of their choice. Speech was recorded on a Zoom H5 using TDK lavalier clip-on microphone. The speech signal was sampled at 44.1 kHz and quantized to 24 bit. In instances where the main microphone failed, backup recordings from the Zoom H2 were used. See [9] for more details. Older participants were aged 40+ and younger participants were aged 16-25. Speakers were all either New Zealand born or arrived in New Zealand under the age of seven. Older speakers must have lived in Auckland for 20+ years.

Table 1. *Vowel Tokens by age and gender*

	Older Women	Older Men	Younger Women	Younger Men
TRAP	666	445	862	713
GOOSE	317	198	377	320
NURSE	280	184	294	252
STRUT	606	325	654	506
START	309	182	354	278
DRESS	677	365	933	775
KIT	723	469	922	688
FLEECE	706	413	650	637
LOT	547	399	611	576
THOUGHT	374	301	409	366
FOOT	150	89	225	200

2.2. Data Preparation

Ten minutes of speech was selected from the 30-minute mark of each interview. This was selected as a time point far enough into the interview for the speaker to be comfortable with the recorder, but before the speaker might be fatigued. The recordings were transcribed using ELAN [13] and passed through WebMAUS (NZ English service) [14]. Further

preparation was done using the EMU-webApp [15]. Phonetic boundaries were hand checked and corrected where necessary. Formant tracks were calculated using EMUR [14] in R [16], then these were hand checked and corrected in the EMU-WebApp where necessary. The vowel targets of stressed monophthongs were labelled based on the criteria given in [17]. The F1 and F2 values were extracted at each vowel target using EMUR in R which was used for the remainder of the analysis. The analysis includes over 20,000 monophthong tokens. Although the focus of this study is the DRESS, TRAP and NURSE vowels, we analyzed all 11 NZE monophthongs to capture the scope of the whole vowel system. *Table 1* provides the number of tokens for each vowel by age and gender.

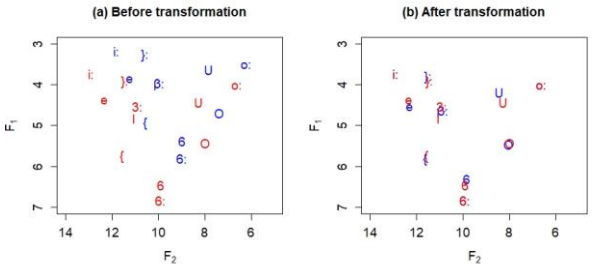


Figure 1. *Raw men's (Blue) and women's (Red) centroids (left), Transformed men's (blue) vowel space on to women's (red) vowel space. (right)*

The statistical analysis includes a data transformation to account for frequency differences between male and female voices. This is a simple linear transformation which transforms male formant values into values closer to female formant using the anchor vowels FLEECE, THOUGHT and START. This results in the transformation shown in *Figure 1*. A full explanation of this transformation is given in [10].

3. Results of Analysis

3.1. Vowel Spaces

Figure 2 shows the results of the formant analysis. The formant values are in bark, with F1 on the Y axis and F2 on the X axis. All four plots show the familiar New Zealand English triangular vowel space, characterized by the centered START and STRUT vowels. All four plots also show a fronted GOOSE vowel, and similar mid-back LOT vowels, as well as raised and retracted FOOT vowels. All groups also have the lowered and retracted KIT vowel distinctive to NZE. A visual inspection of the data, however, suggests that TRAP and DRESS are lowered and retracted for younger men and women in the Auckland Voices database. In addition, NURSE is lowered, patterning with KIT rather than GOOSE for both younger groups. Between the two younger groups there are few visual differences in terms of the overall vowel space, whereas for the older groups the DRESS and NURSE vowel are marginally closer among the older women.

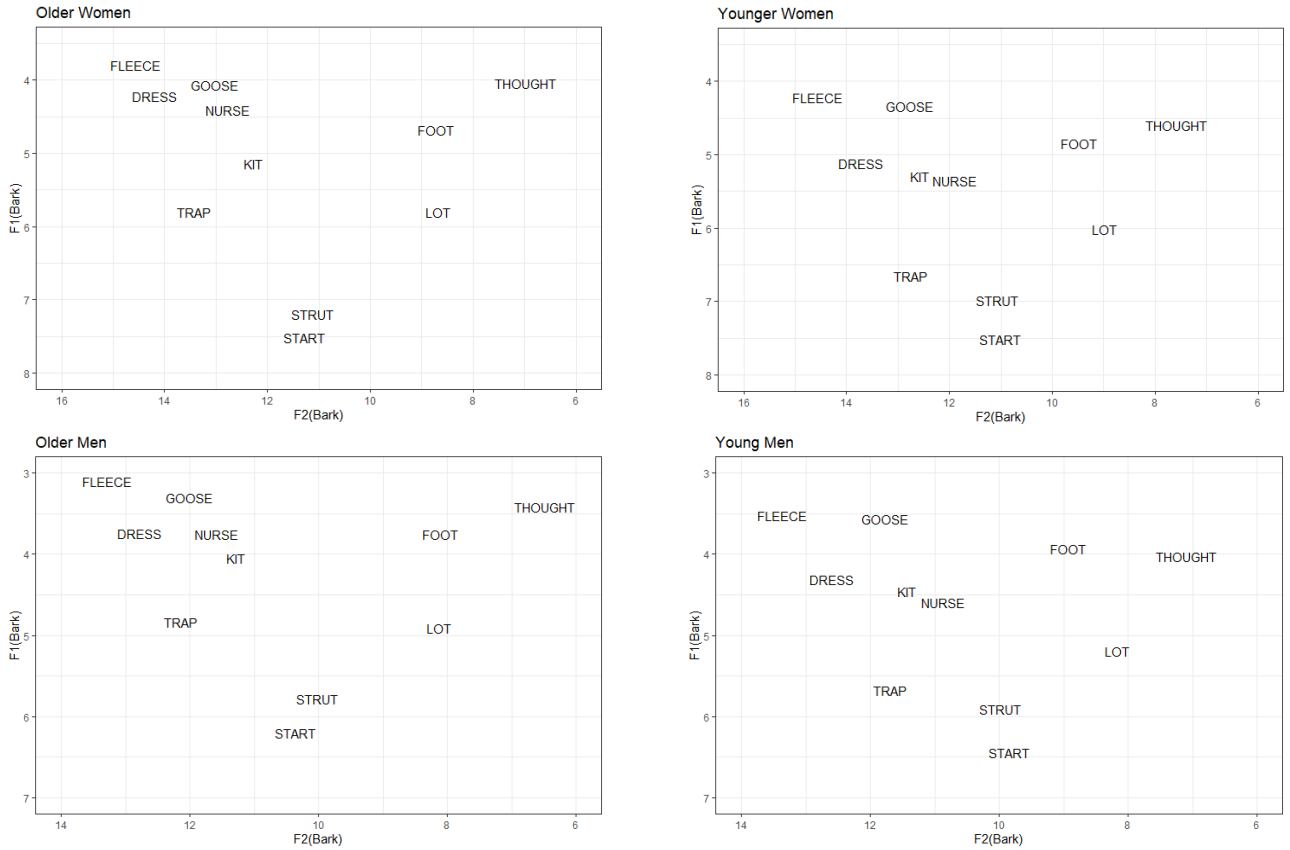


Figure 2: *F1/F2 plots of all speaker groups. Centroids are untransformed means.*

3.2. Statistical analysis

We performed a statistical analysis to investigate whether the differences observed in *Figure 2*. were significant. All statistical analyses were performed in R [16], following the same methodology as [10] and [18]. The linear mixed models were calculated using the `lme()` function in the `nlme` package. This statistical analysis also models F1 and F2 simultaneously in the 2-D formant space. This allows for changes in any direction of the formant plane to be detected. For each vowel four linear mixed models were built for observations of the first and second formants. All models had speaker as a random effect. Fixed effects for each model are given in Table 2. Comparison between models was done with the `nlme` package using the `anova()` function.

Table 2. *The four models used in the investigation.*

Model	*Fixed Effects*	*Random Effects*	*Observation*
Null	Type	Speaker	Formant value
g1	Type*Sex	Speaker	Formant value
g2	Type*Age	Speaker	Formant value

When the null model was compared with g1 (Type*Sex) there were no significant differences at a significance level of 0.01. This suggests that gender differences in this dataset are negligible. When the null model was compared with g2 (Type*Age) there were eight vowels which differed significantly at a significance level of 0.01. These are shown in Table 3. Further analysis from post-hoc t-tests were completed with the older group set as the reference. Full results cannot be reported here for space reasons. The t-test results for our three vowels of interest DRESS, TRAP and NURSE are as follows. Younger speakers DRESS vowel is significantly lower and retracted than for the older speakers (ageY:t (66)=8.14,p<0.01; typeF2:ageY: t(5421)= -7.46,p<0.01). The TRAP vowel is also lowered and retracted for the younger speakers (ageY: t(66)= 9.07,p<0.01; typeF2:ageY: t(5287)= −11.34,p<0.01). Finally, NURSE is also retracted and lowered for the younger speaker group (ageY: t(66)=10.19, p<0.01; typeF2:ageY: t(5287)= -10.68,p<0.01).

Table 3. *Null model vs g2 model – significant differences (significance level 0.01).*

	Degrees of Freedom	AIC Difference	Log Likelihood Ratio	P-value
TRAP	18	68.5	72.502	<.0001
NURSE	18	70.107	74.106	<.0001
START	18	10.686	14.686	6e-04
DRESS	18	43.937	47.937	<.0001
KIT	18	12.08	16.075	3e-04
FLEECE	18	15.141	19.141	1e-04
THOUGHT	18	28.548	32.548	<.0001
FOOT	18	15.282	19.282	1e-04

In addition to the statistical analysis *Table 4.* gives the mean formant values in Bark for the first and second formants for each significant vowel. The mean is calculated from women speakers and the transformed men's values shown in *Figure 1*. Similar observations can be made to those found in the visual and statistical analysis. Most notably, the difference between the F1 and F2 means for DRESS, TRAP, and NURSE when comparing the older and younger speakers.

Table 4. *Mean values of F1 and F2 given in Bark vowel target (T1 and T2) for TRAP, NURSE, START, DRESS, KIT, FLEECE and FOOT by age and gender.*

Vowel	Group	F1	F2
TRAP	Old W.	5.33	11.89
	Old M.	5.25	11.92
	Young Y.	5.99	10.61
	Young M.	6.12	10.7
NURSE	Old W.	4.09	11.38
	Old M.	4.1	11.15
	Young Y.	5.02	9.47
	Young M.	4.94	9.75
START	Old W.	6.86	10.05
	Old M.	6.67	10.44
	Young Y.	6.65	10.2
	Young M.	6.88	9.73
DRESS	Old W.	3.93	12.58
	Old M.	4.1	12.22
	Young Y.	4.89	11.03
	Young M.	4.69	11.52
KIT	Old W.	4.7	10.89
	Old M.	4.42	11.27
	Young Y.	4.62	11.47
	Young M.	4.82	11.19
FLEECE	Old W.	3.56	12.89
	Old M.	3.51	12.85
	Young Y.	3.91	12.42
	Young M.	3.86	12.64
THOUGHT	Old W.	3.8	6.37
	Old M.	3.71	6.42
	Young Y.	4.18	6.62
	Young M.	4.28	6.58
FOOT	Old W.	4.32	7.88
	Old M.	4.04	8.15
	Young Y.	4.23	8.66
	Young M.	4.3	8.66

3.3. Effects of Aging Voices

Although our vowels of interest are significantly different, it should be noted that other vowels show significant difference where we might not expect them to. For example, our anchor vowels FLEECE and THOUGHT. We believe that this is a result of our models not considering the impact of age on the vocal tract. While we performed a transformation to account for gender based vocal tract differences, we did not perform one to account for age based vocal tract differences. Harrington, Palethorpe and Watson [19] look at the impact of aging on formant values and find that as speakers get older their mean F1 and the F2 values, particularly of back vowels, lower. While the data presented in [5] demonstrates the impact aging has on the NZE vowel space specifically.

We can see these effects in place among our data if we look at *Figure 3.* comparing the centroids of our older and younger women. Visual inspection here suggests a similar pattern found in [5, 19]. For example, the centroid for FLEECE is much higher for the older women than the younger women and THOUGHT and FOOT are higher and more retracted. This is

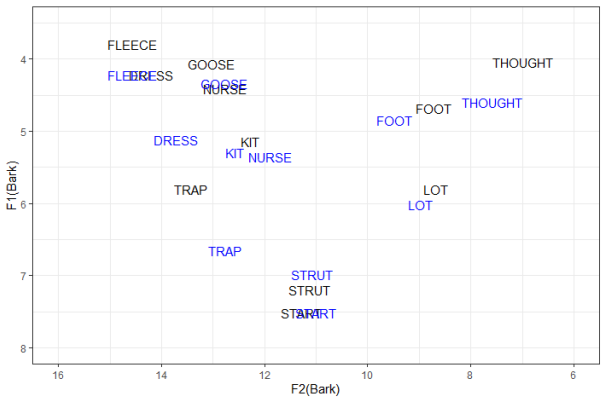


Figure 3. *Centroids of young women (blue) and older women (black).*

an important finding as it suggests the importance of looking at the overall shape of the vowel space rather than means in isolation. To account for this in future analyses, we have several options to consider, such as incorporating aging effects into our linear transformation, using a Vowel Space Measure (VSM) to calculate the distance between vowels in the vowel space such as in [5], or using an alternative method of vowel normalization.

4. Discussion and Conclusion

In this study we compared the speech of two groups of Auckland based NZE speakers. A visual inspection of Figure 1. clearly indicates that for young speakers the DRESS, TRAP and NURSE vowels are lowered and retracted. In particular, for DRESS and NURSE, the centroids of the older women are approaching those of FLEECE and GOOSE respectively, whereas for the younger speaker groups they are more similar in height to the KIT vowel. Additionally, we can conclude that the lowering and retraction may be motivated by the TRAP vowel, as the lack of significant age and gender interactions for KIT suggests it is not undergoing change at this stage. While our statistical analysis supports these findings, it also finds significant differences that might not be expected for example FLEECE, THOUGHT, and FOOT. We believe this is due to aged-based vocal tract differences – not due to phonetic shift. This limits the strength of the findings from the statistical analysis.

Nevertheless, there is an important preliminary finding from this study, which is sound change among NZE speakers in Auckland. If the sociolinguistic patterns of NZE as identified in [20] remain true, this finding is important because it involves the reversal of two well established patterns of change in NZE. Those being short front vowel raising and NURSE raising. In NZE, older women tend to use more conservative variants, and younger speakers, in particular younger women, use more innovative variants [20]. In the case of these Auckland speakers, the younger speakers have lowered and retracted vowels indicating the direction of change is towards these lowered and retracted variants. This contradicts NZE research from outside Auckland which has maintained that vowel raising for DRESS, TRAP and NURSE is ongoing [5, 6]. With DRESS raising to be as high and front as FLEECE, and NURSE raised towards GOOSE. It does, however, support findings from our previous research looking exclusively at the younger speakers in Auckland [9, 10]. It also supports earlier research looking at Auckland based Pasifika speakers of NZE [21].

These findings can be interpreted in two different ways. First, they may suggest there is a difference between the English spoken in Auckland and the English spoken elsewhere in New Zealand. Or alternatively, we could be seeing changes to NZE emerging and spreading from the most linguistically diverse and innovative part of the country. An analysis of a group of NZE speakers from outside of Auckland is underway to provide more insight into these findings. It is also of some interest that the lowering of DRESS and TRAP also mirrors similar research on Australian English [22, 23], which has found lowering for these two vowels amongst Sydney based speakers, a comparably large and diverse city.

Acknowledgements

We thank the NZ Royal Society Marsden fund for supporting the project. We also thank the participants in the Auckland Voices project, and the research assistants involved in recording and transcribing the data.

5. References

[1] Watson, C., Maclagan, M., and Harrington, J., "Acoustic evidence for vowel change in New Zealand English", Language variation and change 12., 12(1): 51-68, 2000.

[2] Maclagan, M., & Hay, J., "The rise and rise of New Zealand English DRESS.", The Proceedings of the Australian International Conference on Speech Science and Technology, 183-188, 2004.

[3] Maclagan, M., & Hay, J., "Getting fed up with our feet: Contrast maintenance and the New Zealand English "short" front vowel shift." Language variation and change, 19(1):1-25, 2007.

[4] Maclagan, M., Watson, C., Harlow, R., King, J., and Keegan, P., "Investigating the sound change in the New Zealand English nurse vowel /ɜ/". Australian Journal of Linguistics, 37(4):465-485, 2017.

[5] Watson, C., Maclagan, M., King, J., Harlow, R., and Keegan, P., "Sound Change in maori and the influence of New Zealand English." Journal of the International Phonetic Association, 46(2):185-218, 2016.

[6] Warren, P., Quality and quantity in New Zealand English vowel contrasts. Journal of the International Phonetic Association, 48(3):305-330, 2018.

[7] Gordon, E., Maclagan, M., & Kortmann, B. (2008). Regional and social differences in New Zealand phonology. Varieties of English, 3, 64-76.

[8] Bauer, L. (1994). "English in New Zealand." In R. Burchfield (Ed.), The Cambridge History of the English Language. Vol5, English in Britain and overseas: origins and development, 382-429.

[9] Ross, B., An acoustic analysis of New Zealand English vowels in Auckland. Master's thesis, Victoria University of Wellington, 2018.

[10] Watson, C., Ross, B., Ballard, E., Charters, H., Arnold, R., & Meyerhoff, M., "Preliminary investigation into sound change in Auckland." In Proceedings of the 17th Australasian International Conference on Speech Science and Technology, 17-20, 2018.

[11] Stats NZ. "Place Summaries | Auckland Region | Stats NZ." https://www.stats.govt.nz/tools/2018-census-place-summaries/auckland-region, accessed on 24 July 2022.

[12] Cheshire, J., Fox, S., Kerswill, P. & Torgersen, E. "Language contact and language change in the multicultural metropolis." Revue Française de Linguistique Appliquée, 17(2):63-76, 2013.

[13] Sloetjes, H., & Wittenburg, P., "Annotation by category – ELAN and ISO DCR." Proceedings of the 6th International Conference on Language Resources and Evaluation, 2008.

[14] Kisler, T., Reichel, U., & Schiel, F., “Multilingual processing of speech via web services.” Computer Speech and Language, 45:326–347, 2017.
[15] Winkelmann, R., Harrington, J., & Jänsch, K., “EMU-SDMS: Advanced speech database management and analysis in R.” Computer Speech and Language, 45:392–410, 2017.
[16] R Core Team, “R: A language and environment for statistical computing.” R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/, accessed on 24 July 2022.
[17] Harrington, J., Phonetic analysis of speech corpora, Wiley Blackwell, 2010.
[18] Watson, C. I., Ballard, E., Ross, B., & Charters, H., “Divergence of FACE and TRAP in Auckland English: A Potential Regional Sound Change in New Zealand English.” The 19th International Congress of Phonetic Sciences, Melbourne, Australia, 2019.
[19] Harrington, J., Palethorpe, S., and Watson, C., "Age-related changes in fundamental frequency and formants: a longitudinal study of four speakers." Interspeech 2007, 8(2), 2007.
[20] Maclagan, M. A., Gordon, E., & Lewis, G., Women and sound change: Conservative and innovative behavior by the same speakers. Language variation and change, 11(1):19-41, 1999.
[21] Starks, D., Gibson, A., & Bell, A., “Pasifika Englishes in New Zealand.” In Williams, J., Schneider, E., Trudgill, p., and Schreier, D., (Eds.), Further studies in the lesser known varieties of English, 288–304, Cambridge University Press, 2015.
[22] Cox, F., & Palethorpe, S., “Reversal of short front vowel raising in Australian English.” Ninth Annual Conference of the International Speech Communication Association, 342-355, 2008.
[23] Grama, J., Travis, C., and Gonzalez, S., "Initiation, progression, and conditioning of the short-front vowel shift in Australia." Proceedings of the 19th International Congress of Phonetic Sciences, 1769-1773, 2019.

An Exploratory Investigation of the /e/-/æ/ and /iː/-/ɪ/ Mergers and Durational Contrasts in Singapore English

Canaan Zengyu Lan, Olga Maxwell, Chloé Diskin-Holdaway

The University of Melbourne

zengyul@student.unimelb.edu.au; omaxwell@unimelb.edu.au;
chloe.diskinholdaway@unimelb.edu.au

Abstract

This study explores the extent of merger between the vowels /e/ and /æ/, /iː/ and /ɪ/, and their durational contrasts, as produced by ten female and male speakers of Singapore English completing a wordlist task. The data were measured acoustically, including F1/F2 values and duration. The results reveal the most overlap for the vowels /iː/ and /ɪ/, followed by /e/-/æ/, /ɪ/-/e/, and /iː/-/e/. The findings also suggest durational differences for vowel pairs, especially among female speakers, whose productions show a greater degree of spectral overlap. This analysis lays the foundation for further investigation examining these vowel contrasts in Singapore English.

Index Terms: Singapore English, vowel merger, durational contrasts, gender differences

1. Introduction

The purpose of this study is to provide new sociophonetic insights into Singapore English (henceforth SgE), which is a variety of English spoken in a highly complex multilingual and multiethnic postcolonial context. The evolution of SgE has been against the backdrop of British colonisation and language contact with local languages such as Malay, Chinese dialects (e.g., Hokkien, Cantonese) and Tamil [1]. There are four official languages in Singapore: English, the interethnic lingua franca and the only medium of instruction in all government schools, and three 'mother tongue' languages (used to denote the speaker's ethnic belonging rather than their L1) — Mandarin, which is the assigned mother tongue for those identifying as ethnically Chinese, Malay for Malays, and Tamil for Indians. Although the 'mother tongues' may not necessarily be spoken at home [2], they have been offered as a single language subject in Singapore since 1987 [3, 4].

Over the years, research on SgE has identified it as an *endonormatively stabilised* postcolonial variety with its own linguistic standards [1, 5]. Such categorisation presupposes linguistic unity and homogeneity in SgE [1, 5]. However, previous work on the acoustic features of vowels in SgE [e.g., 6, 7, 8, 9, 10, 11, 12, 13, 14] suggests that the vowel system of SgE is still emergent, and relatively consistent and predictable pronunciations have yet to be fully established. In particular, there is a tendency for SgE speakers to reduce monophthong contrasts, such as those between front vowel pairs, e.g., /iː/ and /ɪ/, and /e/ and /æ/ [8, 9]. There is a growing body of work on mergers between /e/-/æ/ and /iː/-/ɪ/ in postcolonial "settler" varieties of English [15, 16, 17, 18], such as work on pre-lateral mergers in New Zealand English [17], and pre-lateral merger-in-progress in Australian English spoken in the state of Victoria [18]. However, limited attention has been given to Englishes that have developed as a result of contact with other, often indigenous, languages spoken alongside English (e.g., SgE, Indian English), with even less attention given to sociophonetic characteristics of vowel mergers [19]. Thus, the present study aims to address this gap and bring new insights into language variation in the multilingual, postcolonial society of Singapore.

Early studies on SgE have suggested an acoustic merger between /e/ and /æ/ (e.g., [12]). However, the degree of overlap between the vowels was reported to vary according to contexts, ranging from having a clear /e/-/æ/ contrast in formal contexts to almost no differentiation in casual speech [9, 11, 13]. It has also been noted that /e/ produced in citation form has a range of realisations such as the diphthongal FACE vowel, and anywhere on the continuum between the monophthongal TRAP and DRESS vowels [9, 11], as a result of highly variable F1 and F2 values, e.g., F2 values can vary from a more central position (e.g., 1800Hz) to a very fronted position (e.g., 3000Hz) in the speakers' vowel space [9]. The diphthongal change of /e/ for certain words (e.g., *egg* but not *peg*) was further suggested as evidence for the emergence of a new variety of English among young Singaporeans, whose productions were reported to have diphthongal pronunciation regardless of their ethnic background [9]. These findings suggest interaction between /e/ and other vowel classes with a potential 'redistribution' of vowels in the vowel space, which calls for the examination of co-variation and potential mergers not only for certain vowel pairs but also across the vowel classes.

Mergers between long and short vowels such as /iː/ and /ɪ/ (also tense-lax) have been reported for several postcolonial Englishes spoken in Asia [e.g., 20 for Malaysian English]. While distinctions based on vowel duration have attracted less attention in the literature, duration was suggested as a potential parameter to distinguish spectrally merged vowels such as /iː/ and /ɪ/ in SgE [9, 13], with little variation found in vowel formants for /iː/ and /ɪ/ in SgE across the ethnic groups (i.e., Chinese, Malays and Indians) and contexts [e.g., 7, 8, 9, 10, 11]. Furthermore, the /iː/-/ɪ/ merger in passage reading has been reported for SgE Malay speakers of both genders, with males having a greater degree of overlap as compared to female speakers [13]. However, in a more formal speech style, such as citation form, [13] found a significant difference in mean durational values between /iː/ and /ɪ/. Further, male Malay speakers exhibited a more compact vowel space as compared to female Malay speakers, suggesting possible gender-based differences in vowel productions and the degree of merger [13]. The findings discussed above emphasise the importance of considering the durational parameter in the investigation of vowel mergers [21]. As pointed out by [22], "it would be too premature […]to suggest that two vowel phonemes are merged just because they show overlapping distributions in a two-dimensional vowel space defined by F1 and F2".

1.1. Aims of the study

To date, there has been little examination of durational contrasts in /e/ and /æ/ and /i:/ and /ɪ/ in SgE with reference to gender, and none have explored the possible overlap within and across DRESS-TRAP and FLEECE-KIT vowel pairs. Based on the acoustic-phonetic analysis of the vowels /e, æ, i: ɪ/, this study aims to answer the following questions:

1. To what extent do DRESS and TRAP, and FLEECE and KIT vowels merge acoustically in SgE, i.e., /i:/and /ɪ/, /e/ and /æ/, /ɪ/ and /e/, /i:/ and /e/, /ɪ/ and /æ/, /i:/ and /æ/?
2. Do SgE speakers use duration to differentiate vowel classes in production when the vowel classes are spectrally merged?
3. Is there any effect of gender on the target vowel productions and possible vowel mergers?

2. Method and Materials

2.1. Materials and Procedures

The data analysed for this study were taken from the National Institute of Education Spoken Corpus of English in Asia (NIESCEA) [23] and were used with the approval of the researchers in charge of the corpus. All audio files were in .wav format and were produced by 10 native speakers of SgE: five females and five males. They were all undergraduate or graduate students, aged between 18 and 35 years at the time of the recording. The 10 audio files included 33 monosyllabic sample words that contained 11 monophthongs. All speakers were asked to read a list of words only once. In this study, two pairs of monosyllabic words were selected which included the front vowels /e, æ/ and /i:, ɪ/. The examined pairs were produced in /hVd/, /bVt/ and /bVd/ contexts with the words embedded in the carrier sentence 'Please say ____ again'. The word set used in this study is included below, with the target words shown in bold font:

Please say **head** again. Please say **had** again.
Please say **bet** again. Please say **bat** again.
Please say **bed** again. Please say **bad** again.
Please say **heed** again. Please say **hid** again.
Please say **beat** again. Please say **bit** again.
Please say **bead** again. Please say **bid** again.

A total of 120 tokens were elicited from 12 citation words from 10 speakers. The words in their respective pairs formed minimal pairs, however, these minimal pairs were randomised and produced separately by all speakers during the recording [23]. The target vowels in their carrier phrases were extracted from the original audio files and saved separately in .wav format for each individual speaker.

2.2. Measurement and analysis procedure

All segmentations of the target vowels were annotated manually in Praat (version 6.1.40) [24]. The vowel duration (in ms) was measured from the left boundary to the right boundary of each vowel. The left boundary of each vowel was measured from the onset of clear formant energy, indicated by regular F1-F4 spectral energy of each vowel, while the right boundary was placed at the offset of spectral discontinuity of each vowel, indicated by a clear change in the overall shape of the vocal tract. A Praat script was used to extract durational values and F1/F2 values at vowel midpoints. Outliers caused by a formant tracking error, such as /i:, ɪ/ with low F2 values, were adjusted manually with reference to the wideband spectrograms, by placing the cursor at the temporal midpoint and in the middle of F1 and F2. The same principle of measurement was followed consistently throughout the measurements.

F1 and F2 characteristics (in Hz) were analysed using the *emuR* package in *R* (version 4.0.4) [25, 26]. *R* libraries such as *ggplot2* and *phonTools* were used to plot F1/F2 values of the target vowels with ellipses and duration in boxplots according to gender. F1/F2 values were normalised using Lobanov normalisation for comparison between the genders, given the effectiveness of this method in preserving the phonemic and the sociolinguistic variation while minimising the gender-related physiological variation [27]. The Pillai-Bartlett Trace, otherwise known as the Pillai score and as an output of a MANOVA, was calculated in *R* with the *tidyverse* package [28] to measure acoustic merger between six pairs of vowels, namely, /i:/and /ɪ/, /e/ and /æ/, /ɪ/ and /e/, /i:/ and /e/, /ɪ/ and /æ/, /i:/ and /æ/. The Pillai score was chosen due to its ability to compare between-speaker differences, capture the degree of overlap between vowel categories in the acoustic space, and conduct analysis on a small number of tokens [29]. The Pillai score values range from 0 to 1, with figures approaching 1 indicating a high level of distinction and values closer to 0 indicating a greater degree of overlap in the F1/F2 space. The significance values (Pr>(F)), as part of the output, were also extracted and presented in parenthesis with the Pillai scores. Mean duration, F1 and F2 values and their standard deviations were calculated according to the adjusted measurement values (the output of the Praat script) in Numbers (version 11.1) according to each vowel and gender category.

3. Results

3.1. Acoustic results

Table 1 presents descriptive statistics and summarises the results of the mean F1 and F2 values (in Hz) and mean duration of /e, æ, i: ɪ/ (in ms) for female and male speakers. For both gender groups, mean F1 values are higher for /æ/ (female - $\bar{x}$ 780Hz, σ 71Hz; male - $\bar{x}$ 615Hz, σ 43Hz), compared to /e/ (female - $\bar{x}$ 617, σ 155Hz; male - $\bar{x}$ 492Hz, σ 102Hz), and mean F2 values are lower for /æ/ (female - $\bar{x}$ 1708Hz, σ 412Hz; male - $\bar{x}$ 1880Hz, σ 93Hz), compared to /e/ (female - $\bar{x}$ 1905Hz, σ 364Hz; male - $\bar{x}$ 2027Hz, σ 185Hz).

Table 1. *Mean F1/F2 values for /e, æ, i: ɪ/ in Hz and mean vowel duration in ms with standard deviation in parenthesis for both female (F) and male (M) speakers.*

	Phoneme	F1 (Hz)	F2 (Hz)	Duration (ms)
F	æ	780(71)	1708(412)	166(101)
	e	617(155)	1905(364)	143(68)
	i:	428(48)	2590(540)	172(88)
	ɪ	467(65)	2172(606)	124(75)
M	æ	615(43)	1880(93)	139(43)
	e	492(102)	2027(185)	129(37)
	i:	320(40)	2214(196)	134(34)
	ɪ	359(40)	2078(109)	109(30)

The results also indicate that mean F1 values of /e/ and /æ/ produced by male speakers are lower than those produced by female speakers, but mean F2 values of /e/ and /æ/ produced by male speakers are higher than those produced by female speakers. The difference in mean F1 values between /e/ and /æ/

produced by male speakers is smaller than that produced by female speakers (123Hz versus 163Hz), and the difference in mean F2 values between /e/ and /æ/ produced by male speakers is also smaller than that of the female group (147Hz versus 197Hz).

Furthermore, for both female and male speakers, mean F1 values are higher for /ɪ/ (female - $\bar{x}$ 467Hz, σ 65Hz; male - $\bar{x}$ 359Hz, σ 40Hz), compared to /i:/ (female - $\bar{x}$ 428, σ 48Hz; male - $\bar{x}$ 320Hz, σ 40Hz), and mean F2 values are lower in /ɪ/ (female - $\bar{x}$ 2172Hz, σ 606Hz; male - $\bar{x}$ 2078Hz, σ 109Hz), compared to /i:/ (female - $\bar{x}$ 2590Hz, σ 540Hz; male - $\bar{x}$ 2214Hz, σ 196Hz). The results also indicate that both mean F1 and F2 values of /i:/ and /ɪ/ produced by male speakers are lower than those produced by female speakers. The difference in mean F1 values between /i:/ and /ɪ/ produced by male speakers is the same as that produced by female speakers (39Hz), while the difference in mean F2 values between /i:/ and /ɪ/ for male speakers is smaller than that for females (136Hz versus 418Hz).

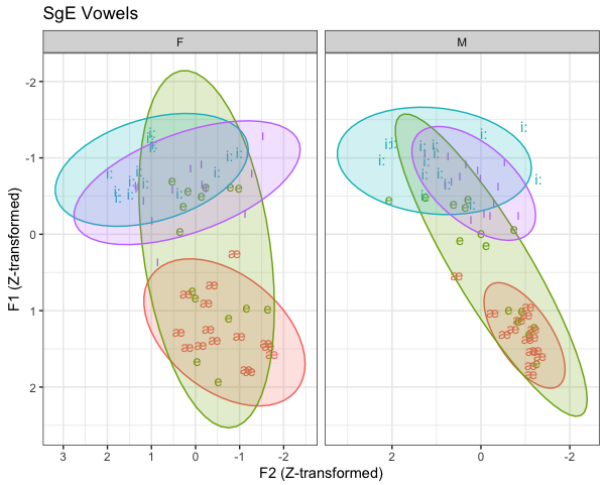


Figure 1: *First and second formant frequency (Lobanov-normalised) at midpoints of /e/ (green ellipsis), /æ/ (red ellipsis), /ɪ/ (purple ellipsis), /i:/ (blue ellipsis) according to gender (female - left, male - right).*

Figure 1 shows two Lobanov-normalised plots of F1 values (*y*-axis) against F2 values (*x*-axis) measured at the vowel midpoints, with ellipses over each vowel production, presented by gender (female speakers – left panel, male speakers – right panel). For both groups, F1 and F2 values for /e/ and /æ/, and /i:/ and /ɪ/ are indicative of the acoustic overlap between not only /e/ and /æ/, and /i:/ and /ɪ/, but also /e/ with both /i:/ and /ɪ/. Moreover, both female and male speaker groups show similar variation patterns, but with female speakers having a more dispersed vowel space than that of male speakers. Further, a closer look at the data shows that the raw F1 values for /e/ vary from 400Hz to 900 Hz for female speakers and 400HZ to 700Hz for male speakers. The raw F2 values for /i:/ vary the most, ranging from 1800Hz to 2500Hz among male speakers, while female speakers have the most variation in the productions of /ɪ/, ranging from 1000Hz to 3000Hz.

Pillai scores used to examine the extent of acoustic overlap between the two vowels in each pair are presented in Table 2, with significance values (Pr>(F)) included in parenthesis. The overall (averaged across all speakers) Pillai scores suggest overlap for a number of vowel pairs: /i:/and /ɪ/ (0.22), /e/ and /æ/ (0.25), /ɪ/ and /e/ (0.29), while vowel pairs such as /i:/ and /e/ (0.49), /ɪ/ and /æ/ (0.73), and /i:/ and /æ/ (0.81) indicate various degrees of distinction. Furthermore, *p*-values for all vowel pairs are below 0.001. When looking at the data combined across the speakers, the patterns suggest that SgE speakers are producing mergers for pairs /i:/ - /ɪ/, /e/ - /æ/, /ɪ/ - /e/ and are making better distinctions for vowel pairs /ɪ/-/æ/ and /i:/-/æ/, and most likely /i:/ - /e/.

Table 2. *Pillai scores for all possible vowel pairs, presented from the lowest to the highest score with significance values (Pr>(F)) in parenthesis.*

	Female	Male	Overall
/i:/&/ɪ/	0.32 (p<0.01)	0.27(p<0.05)	0.22 (p<0.001)
/e/&/æ/	0.33 (p<0.01)	0.43 (p<0.001)	0.25 (p<0.001)
/ɪ/&/e/	0.34 (p<0.01)	0.59 (p<0.001)	0.29 (p<0.001)
/i:/&/e/	0.55 (p<0.001)	0.57 (p<0.001)	0.49 (p<0.001)
/ɪ/&/æ/	0.86 (p<0.001)	0.92 (p<0.001)	0.73 (p<0.001)
/i:/&/æ/	0.91 (p<0.001)	0.93 (p<0.001)	0.81 (p<0.001)

When examined on the basis of gender, Pillai scores indicate that male speakers have a marginal difference in the degree of overlap between /i:/ and /ɪ/ (0.27) compared to female speakers (0.32). However, female speakers produce a greater amount of overlap between /e/ and /æ/ (0.33), as compared to male speakers who produced a more modest overlap in this pair (0.43). In addition, although the overlap between the vowels /ɪ/ and /e/, /i:/ and /e/ is more partial for both male and female speakers, female speakers tend to have greater amount of overlap for the /ɪ/ and /e/ vowel pair (0.34) compared to the /i:/ and /e/ vowel pair (0.55). This is in contrast to male speakers whose patterns show that /e/ overlapped more with /i:/ (0.57) and /ɪ/ (0.59). Lastly, the Pillai statistic results suggest that both gender groups make a distinction between /ɪ/ and /æ/, /i:/and /æ/, with the scores of 0.92 and 0.93 for male and 0.86 and 0.91 for female speakers. The distinction between /ɪ/ and /æ/, /i:/and /æ/ is also significant for both genders (p<0.001).

3.2. Duration results

Due to the small sample size, t-test results of durational contrasts did not show any significant differences between the examined vowel pairs (except for /i:/-/ɪ/ which has p<0.05), thus, *p*-values are not presented in this section. Figure 2 illustrates the durational values of vowels presented by gender, with the female group on the left and the male group on the right. Duration in ms is plotted on the y-axis and the target phonemes on the x-axis. For both female and male speaker groups, /æ/ is longer than /e/ and /i:/ is longer than /ɪ/, according to their median values as indicated by the horizontal line in each box. The durational differences between /e, æ, i: ɪ/ are very small, according to their median values. In general, male speakers produce slightly longer vowels across two pairs, as indicated by higher median values. Female speakers show greater variability in production than male speakers, as indicated by the shape of the boxes, more extended whiskers, and the presence of more outliers. /i:/ has the most variability in duration for female speakers (from 110ms to 210ms), but it is the least variable vowel for male speakers (from 110ms to 150ms). /e, æ, ɪ/ vary to almost the same degree for female speakers (80ms difference), but for male speakers, /æ/ (60ms difference) varies more than /e/ (50ms difference) which in turn, varies more than /ɪ/ (40ms difference).

Similarly, the mean duration values shown in Table 2 suggest that for both female and male speakers, /æ/ (female - $\bar{x}$ 166ms, σ 101ms; male - $\bar{x}$ 139ms, σ 43ms) is longer than /e/ (female - $\bar{x}$ 143ms, σ 68ms; male - $\bar{x}$ 129ms, σ 37ms) and /i:/

(female - x̄ 172ms, σ 88ms; male - x̄ 134ms, σ 34ms) is longer than /ɪ/ (female - x̄ 124ms, σ 75ms; male - x̄ 109ms, σ 30ms). However, in contrast to median values shown in Figure 2, the mean duration results indicate that females produced longer vowels for both pairs compared to male speakers. /iː/ is still the longest vowel produced by female speakers but not for male speakers, whose longest vowel is /æ/. Such a finding could be the result of averaging the amount of variability within the vowel classes where female speakers produce the most variation in /iː/ while male speakers have the most variation in /æ/, indicated by their standard deviations. Female speakers also produce larger durational differences in /e, æ/ (23ms), /iː, ɪ/ (48ms) and /iː, e/ (29ms) but smaller in /ɪ, e/ (19ms) than male speakers (10ms for /e, æ/; 25ms for /iː, ɪ/; 5ms for /iː, e/ and 20ms for /ɪ, e/).

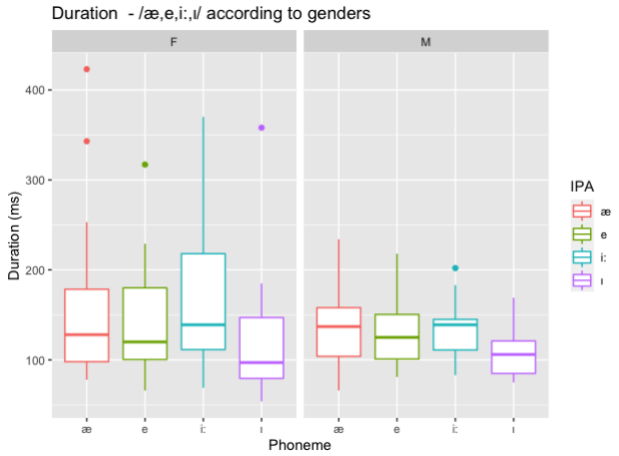


Figure 2: *Duration for /e, æ, ɪ, iː/ according to gender (female - left, male - right).*

We also observed that some speakers tend to have longer stops before target words (e.g., after *please say*), followed by a strong emphasis on the target words, resulting in exceptionally long vowel productions and hyperarticulation (e.g., one male and one female speaker were found to use 3s and 4s respectively to produce *hid* in the carrier phrase, as compared to 1s by most other speakers). As a consequence, some of the variation may be attributed to the nature of the task and the way some participants produced the vowels.

4. Discussion and conclusion

While the acoustic results indicate evidence of merger behaviour between /e/ and /æ/, and between /iː/ and /ɪ/, the patterns of such mergers are gradient for both female and male speakers. Such results, to some extent, corroborate findings reported in previous studies [8, 9, 11, 13] that these two vowel pairs tend to overlap and that there is indeterminate intra-speaker variation. However, the results of the present study suggest that such overlap for wordlist data is partial, especially for /e/ and /æ/ among male speakers. The findings also reveal that /ɪ/ and /e/ overlap for female speakers, but only partially so for male speakers; similarly, /iː/ and /e/ also overlap, although partially for both genders. Furthermore, the current results corroborate /e/-raising, suggested in previous work on SgE [e.g., 10, 11, 14].

In addition, the results of the current study show that there is greater internal variation and more dispersed productions among female speakers than among male speakers. This is consistent with the study by [13] and confirms their observations of overlap patterns based on gender, where female speakers have a greater degree of overlap for /e/ and /æ/, but smaller for /iː/ and /ɪ/. In contrast to the findings reported in other studies on SgE (e.g., [13]), we did not find a more fronted pronunciation of /æ/ as compared to /e/ among male speakers. Although /iː/ and /ɪ/ are the most overlapped pair for both genders, their durational difference is also the largest and highly perceivable for both genders (e.g., >20ms).

In general, female speakers have a greater degree of overlap for all pairs, except for /iː/ and /ɪ/; at the same time, female speakers produce greater and more contrastive and perceivable durational differences for all pairs (>20ms except for /ɪ, e/ of 19ms). In comparison, vowels produced by males overlap less and the durational contrasts are also smaller, with only two potentially perceivable pairs, /ɪ, e/ (20ms) and /iː, ɪ/ (25ms). Such observations suggest that there seems to be a complementary relationship between vowel quality and vowel duration. As such, it could be inferred that speakers of SgE, especially female speakers, may rely on durational differences to distinguish vowel differences in speech. However, as suggested in previous instrumental work (e.g., [12]), the quality of SgE vowels vary across contexts, with a greater likelihood of mergers in casual speech and greater distinctions in formal contexts. This study investigated vowels in citation form, which is a more formal and carefully controlled method as compared to casual conversations as well as the passage and the sentence reading tasks. Thus, it could be predicted that such durational contrasts might not appear in informal everyday conversations. Further studies with a larger sample and comparisons between groups and contexts are required for a more robust conclusion.

Furthermore, there is a lot of variation in the F2 values for the speakers in this study, consistent with other investigations of SgE [e.g., 9]. Careful inspection of the data indicate that one male speaker and one female speaker tended to produce more backed close vowels than other speakers. Hence, F2 values could have been affected by speaker-intrinsic variation in our sample. Moreover, each speaker was asked to read the words in citation form only once, limiting the number of tokens. Another possible reason for such variation in F2 could be due to the differences in pronunciation across the three ethnic groups in Singapore. Recall that the NIESCEA does not provide details about the speakers' linguistic, ethnic, and social background. As suggested in the literature [e.g., 9, 11, 13], Indian Singaporean speakers tend to maintain a clear /e/-/æ/ contrast whereas Malay Singaporeans produce the least distinction. Therefore, we emphasise the importance of collecting speakers' ethnic and language background details in future studies.

Given its exploratory nature and a small sample based on citation form data, the conclusions rendered in this study are tentative and subject to further investigation. Nonetheless, this exploratory study helps to address the gap in the recent literature on mergers in SgE. Our future work will include the collection of a large corpus of data based on present-day SgE speech, with a focus on different speech types (e.g., natural versus formal speech) and speakers' ethnic backgrounds. Further, a more nuanced approach to social factors such as gender will be taken where speakers can self-identify their gender in order to provide a more comprehensive picture of the vowel productions and mergers in SgE.

5. References

[1] Schnerider, E.W., Postcolonial English: Varieties Around the World". Cambridge: Cambridge University Press, 2007.

[2] Dixon, L. Q., "Assumptions behind Singapore's language-in-education policy: implications for language planning and second language acquisition", Lang Policy 8, 117–137, 2009.
[3] Alsagoff, L., "English in Singapore: Culture, capital and identity in linguistic variation: English in Singapore: culture, capital and identity in linguistic variation", World Englishes, 29(3), 336–348, 2010.
[4] Census of 2020, "Census of Population 2020 Statistical Release 1: Demographic Characteristics, Education, Language and Religion", Depart of Statistics Singapore, 2020.
[5] Schneider, E. W., "Models of English in the world", in M. Filppula, J. Klemola, D. Sharma [Eds], The Oxford handbook of world Englishes, 35-57, Oxford University Press, 2017.
[6] Deterding, D., "Measurements of the /eɪ/ and /əʊ/ vowels of young English speakers in Singapore", in A. Brown, D. Deterding, and E.L. Low [Eds], The English Language in Singapore: research on Pronunciation, 93–99, Singapore: Singapore Association for Applied Linguistics, 2000.
[7] Deterding, D., "The measurement of rhythm: A comparison of Singapore and British English", Journal of Phonetics, 29, 217–230, 2001.
[8] Deterding, D., "An instrumental study of the monophthong vowels of Singapore English", English World-Wide, 24, 1–16, 2003.
[9] Deterding, D., "Emergent patterns in the vowels of Singapore English", English World-Wide, 26, 179–197, 2005.
[10] Deterding, D., "Phonetics and Phonology". In D. Deterding [Ed], Singapore English, 12-39, Edinburgh University Press, 2007.
[11] Deterding, D., "The vowels of the different ethnic groups in Singapore", In D. Prescott, A. Kirkpatrick, I. Martin, A. Hashim [Eds], English in Southeast Asia: Literacies, Literatures and Varieties, 2–29, Newcastle, UK: Cambrdge Scholars Press, 2007.
[12] Suzanna, B. H. and Adam, B., "The [e] and [æ] vowels in Singapore English". In A. Brown, D. Deterding, and E. L. Low [Eeds], The English Language in Singapore: research on Pronunciation, 84–92, Singapore: Singapore Association for Applied Linguistics, 2000.
[13] Tan, R. and Low, E., "How different are the monophthongs of Malay speakers of Malaysian and Singapore English", English World-wide, 31, 162-189, 2010.
[14] Tay, W. J. Mary, "The phonology of educated Singapore English", English World-Wide, 3, 135-145, 1982.
[15] Schmidt, P., Diskin-Holdaway, C. and Loakes, D., "New insights into /el/-/æl/ merging in Australian English", Australian Journal of Linguistics, 41(1): 66–95, 2021.
[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", Australian Journal of Linguistics, 34(4):436–452, 2014.
[17] Hay, J., Drager, K. and Thomas, B., "Using nonsense words to investigate vowel merger", English Language and Linguistics, 17(2): 241–269. https://doi.org/10.1017/S1360674313000026, 2013.
[18] Diskin-Holdaway, C., Loakes, D., Billington, R., Stoakes, H. and Gonzalez, S., "The /el-/ael/ merger in Australian English: Acoustic and articulatory insights", Proceedings of the International Congress for the Phonetic Sciences, Australasian International Conference on Speech Science & Technology, Melbourne, 2019.
[19] Gu, Y. and Chen, N. F., "Large-Scale Acoustic Characterization of Singaporean Children's English Pronunciation", 2022.
[20] Pillai, S., Mohd. Don, Z., Knowles, G. and Tang, J., "Malaysian English: an instrumental analysis of vowel contrasts", World Englishes, 29(2):159–172, 2010.
[21] Wade, L., "The role of duration in the perception of vowel merger", Laboratory Phonology: Journal of the Association for Laboratory Phonology, 8(1):1–34, 2017.
[22] Gordon, M. J., "Investigating chain shifts and mergers", in J. K. Chambersand, N. Schilling, N., [Eds], The Handbook of Language Variation and Change, 205, JohnWiley & Sons, 2013.
[23] Low, E. L., "The NIE Spoken Corpus of English in Asia (NIESCEA)", Singapore: National Institute of Education, Nanyang Technological University, 2015.
[24] Boersma, P. and Weenink, D., Praat: Doing phonetics by computer [Version 6.1.40], Available: http://www.praat.org/, 2021
[25] R Core Team, R: A language and environment for statistical computing [Version 4.2.1]. Available: https://www.r-project.org, 2022.
[26] Winkelmann, R., Jänsch, K., Cassidy, S., and Harrington, J., emuR: Main package of the EMU Speech Database Management System [Version 1.4.1]. Available: https://cran.r-project.org, 2021.
[27] 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.
[28] Wickham, H., Mara, A., Jennifer, B., Winston, C., Lucy, M., Romain, F.,...and Hiroaki, Y., "Welcome to the Tidyverse", Journal of Open Source Software, 4(43):1686–1671, 2019.
[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.

A corpus-based computational analysis of high-front and -back vowel production of L1-Japanese learners of English and L1-English speakers

Martin Schweinberger, Yuki Komiya

The University of Queensland

m.schweinberger@uq.edu.au, y.komiya@uq.net.au

Abstract

This study combines acoustic phonetics with computational and applied corpus linguistics to analyze and compare the production of the monophthongal vowels /ɪ/, /iː/, /ʊ/, and /uː/ in the speech of 148 L1-Japanese learners (JPN) and 107 L1-speakers of English (ENS) based on *The International Corpus Network of Asian Learners of English* (ICNALE). The study aims to ascertain if JPN merge spectrally close vowels by calculating Bhattacharya coefficients. In addition, the study uses mixed-effects linear regression to determine if JPN compensate for the potential mergers by exaggerating durational contrasts between spectrally similar vowels. The results of the analysis confirm that JPN exhibit high degrees of overlap for both /ɪ iː/ and /ʊ uː/. Their L1-English peers, however, also exhibit substantive overlap for /ʊ uː/. With respect to duration, the analysis shows that JPN extend the duration of all vowels and exaggerate the difference between /ɪ iː/ and /ʊ uː/ to compensate for the lack of qualitative differences between short and long vowel pairs. This study represents the first corpus-based acoustic analysis of JPN vowels in spontaneous speech.

Index Terms: acoustic phonetics, vowel production, learner corpus research, Japanese learners of English

1. Introduction

While pronunciation poses a challenge for language learners, it is also the most immediate and direct display of linguistic proficiency. Listeners automatically and subconsciously categorize and infer judgments about speakers based on pronunciation [2]. In addition, pronunciation is crucial for intelligibility and is affecting real-life opportunities (jobs, partner choice, etc.).

A underlying cause for the difficulties that learners face is that languages are not independent but interact in the minds of multilingual speakers [2] which means that the L2 sound system is affected by the L1 system (and vice versa). From the perspective of JPN, English vowels are particularly challenging [3] due to

- Differences in inventory size (Japanese: 5 monophthongal vowels vs. English: app. 11 monophthongal vowels (depending on the variety of English)) [4]
- Differences in how vowels are differentiated (Japanese: duration differences versus English: formant *and* duration differences)

Formants are concentration of acoustic energy at a certain frequency [5] with the first formant (F1) and the second formants (F2) of a vowel sound inversely corresponding to the tongue height and tongue fronting during vowel production. Regarding the production of English vowels produced by JPN, it has been shown that JPN merge spectrally similar vowels (including high-front and -back vowels) [6]. Furthermore, it has been reported that JPN are very sensitive to vowel duration [7] and exaggerate duration to compensate for the relative insensitivity to formant differences [8].

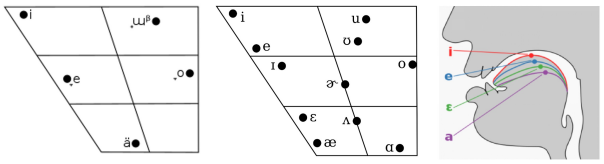


Figure 1: *left panel: vowel chart showing monophongal vowels of standard Japanese; center panel: vowel chart of Southern Californian American English; right panel: tongue position corresponding to selected front vowels.*

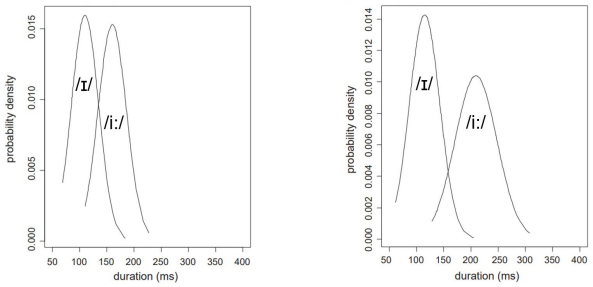


Figure 2: *left panel: durations of high front vowels produced by ENS; right panel: durations of high front vowels produced by JPN.*

Previous research on vowel production by JPN is predominantly based on read-aloud word lists or selected scripted sentences in highly controlled laboratory conditions. Hence, characteristics of the vowel production of learners in naturalistic speech environments remain largely unknown. Furthermore, previous research has relied on small samples of subjects with studies using between 8 and 15 subjects. As such, the findings provided by previous research may not warrant generalization to larger speech communities or to conversational language production in natural settings.

The present study addresses these issues and aims to provide a more detailed understanding of the following research questions:

1. Do JPN merge /iː/ and /ɪ/ as well as /uː/ and /ʊ/?
2. Do JPN exaggerate the lengths of vowels to compensate for a lack of spectral differentiation?

2. Corpus Data

The study uses data from the *International Corpus Network of Asian Learners of English* (ICNALE) [9]. The ICNALE is one of the largest publicly available multimodal learner corpora comprising more than 10,000 topic-controlled speeches and essays produced by college students in ten countries and regions in Asia as well as English native speakers. For this study, all data representing spoken monologues (spontaneous speech) collected between 2017 and 2019 from 148 JPN and 107 ENS were analyzed (this encompasses all JPN and ENS speakers in the data that produced relevant tokens and, in the case of ENS, spoke American English).

Every speaker contributed two one-minute recording to the spoken monologues' component of ICNALE. Speech samples were recorded on mobile devices or personal computers using in-built microphones resulting in a highly variable quality of recordings.

3. Data Processing

Data processing started with using Web-MAUS [10] to (force) align the audio files and transcriptions provided by ICNALE into Praat TextGrids (the forced alignment used both US and British models). All subsequent steps of the analysis were performed using R Version 4.2 [11] in RStudio [12].

The first to third formants of all vowels as well as vowel durations were extracted using rPraat [13], wrassp [14], and tidyverse [15]. The algorithm targeted a range between three and 7 formants for each vowel resulting in five formant values for each of the first to third formants in each vowel. The optimal formant values out of these five options were determined based on the minimal Euclidean distance to standard American English vowel formants based on [16] and standard southern British English based on [17]. Based on the Euclidean distance, alternative options and the less well fitting target variety were removed from the analysis so that the data set contained only one observation for each vowel (see Table 1)

Table 1: *Overview of the semi-processed data set.*

Type	Speakers	/ɪ/	/iː/	/ʊ/	/uː/	Total
ENS	132	1,077	1,182	348	513	3,120
JPN	149	2,779	1,084	608	623	5,094
Total	281	3,856	2,266	956	1,136	8,214

In a next step, socio-demographic information about the speakers (speaker type, age, gender, English proficiency) was added to the data and ENS not from North America were removed. All further analyses continued with standard American English as target variety. Table 2 provides an overview of the socio-demographics of the speakers in the data and Table 3 shows the proficiency levels among the L1 Japanese learners.

Next, vowels were normalized using a z-transformation after grouping the data by speaker type (ENS vs JPN) and gender. After this normalization procedure, all vowels not representing /iː/, /ɪ/, /uː/, and /ʊ/ were removed from the analysis.

Then, multi-syllabic words or words containing more than 9 characters were removed to better control for variability caused by the phonetic and phonological environments in which the vowels were produced. Only words were retained which had a /CV(C)/ syllable structure (e.g., *get*, *gut*, *hit*, *shit*, *due*, *we*, *see*).

To account for the low quality of audio recordings and to remove outliers and inaccuracies, kernel density estimation was applied to the z-transformed first and second formats. All vowels having density values in the lower quartile of first and second formats were removed. The final data set is summarized in Table 4.

Table 2: *Overview of information about the speakers represented in the data.*

Type	Gender	18-29	30-39	40-49	50+	Total
ENS	female	25	5	1	4	35
	male	30	27	9	6	72
JPN	female	63	1	0	0	64
	male	82	1	1	0	84
Total		200	34	11	10	255

Table 3: *Overview of proficiency levels among JPN speakers in the data.*

Gender	A2	B1	B2	Total
female	14	38	12	64
male	16	50	18	84
Total	30	88	30	148

Table 4: *Overview of the final data set (percentage of retained observations compared to semi-processed data in brackets).*

Type	Speakers	/ɪ/	/iː/	/ʊ/	/uː/	Total
ENS	107	560	785	159	311	1,815
	(81.0)	(52.0)	(66.4)	(45.7)	(60.6)	(58.2)
JPN	148	917	481	155	231	1,721
	(99.3)	(33.0)	(44.4)	(25.5)	(37.1)	(33.6)
Total	255	1,477	1,203	314	542	3,536
	(90.7)	(38.3)	(53.1)	(32.8)	(47.7)	(43.0)

4. Statistical Analysis

The statistical analysis made use of two procedures

- Bhattacharya coefficients: to assess potential spectral mergers of /iː/ and /ɪ/ as well as /uː/ and /ʊ/
- Mixed-effects linear regression: to assess if JPN exaggerate the length of vowels to compensate for a potential lack of spectral differentiation

Bhattacharya coefficients are suited to assess vowel mergers as this coefficient represents a measure of overlap of scatter clouds with 1 representing perfect overlap and 0 representing zero overlap.

Mixed-effects linear regression modeling was performed using the lme4 [18] and the sjPlot package [19] with a step-wise step-up model fitting procedure. The regression analysis evaluated the effect of the following variables and their two-way interactions. If models exhibited substantial multicollinearity (variance inflation factors ≤ 5, the model was considered not trustworthy). Table 5 details the variables that were tested during the statistical modeling and provides information about their scaling as well as how they were operationalized.

Proficiency could not be included into the model and used as a predictor as no proficiency information was available for ENS. Including proficiency as a predictor would have led to the

Table 5: *Variables included in the mixed-effects regression modelling (ran. eff. = random effect, ind. var. = independent variable, cat. = categorical scaling, nom. = nominal scaling).*

Variable	Type	Scale	Levels / Description
duration	dep. var.	num.	duration of vowel in ms
speaker	ran. eff.	cat.	id of speaker
word	ran. eff.	cat.	word, e.g., *good*, *foot*,*he*, *hit*
vowel	ind. var.	cat.	vowel sound: /iː/, /ɪ/, /uː/, /ʊ/
type	ind. var.	nom.	vowel produced by ENS or JPN
status	ind. var.	nom.	status of word (gramm. vs lex.)
gender	ind. var.	nom.	female vs. male (self-reported)
age	ind. var.	num.	age of speaker in years

automatic exclusion of all ENS data points or it would have resulted in ill-fit models due to the absence of variability in proficiency among ENS.

5. Results

The following reports on the findings of the analysis separated by research question the statistical analyses have addressed.

5.1. Vowel mergers

Regarding the potential merger between /iː/ and /ɪ/ among JPN and ENS speaker, the high values of the Bhattacharya coefficient confirmed the expected merger of spectrally similar high-front vowels among JPN but not among ENS:

- JPN Bhattacharya coefficient (/iː/, /ɪ/): .870
- ENS Bhattacharya coefficient (/iː/, /ɪ/): .668

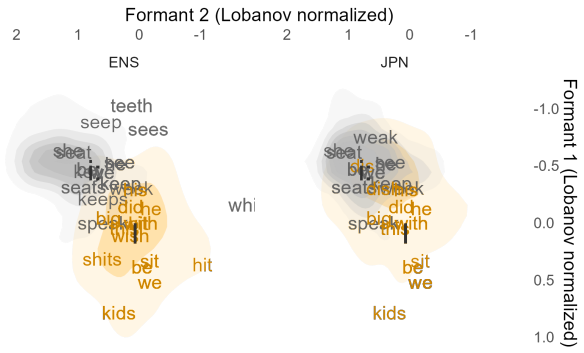


Figure 3: *left panel: overlap of high-front vowels among JPN; right panel: overlap of high-front vowels among ENS.*

Regarding the potential merger between /uː/ and /ʊ/ among JPN and ENS speaker, the high values of the Bhattacharya coefficient confirmed the expected merger of spectrally similar high-back vowels for JPN but also confirmed a merger of spectrally similar high-back vowels among ENS:

- JPN Bhattacharya coefficient (/uː/, /ʊ/): .881
- ENS Bhattacharya coefficient (/uː/, /ʊ/): .929

Thus, the analysis confirmed the expected mergers of spectrally similar high-front and -back vowels among JPN but also a merger of high-back vowels among ENS.

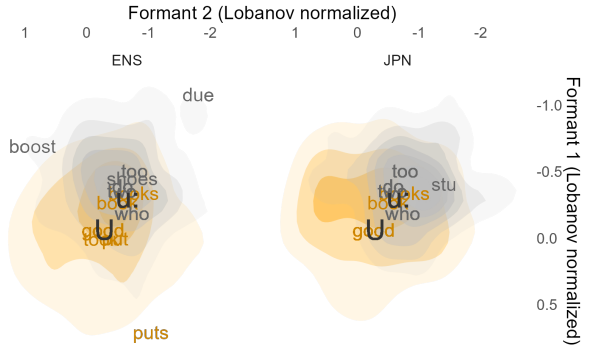


Figure 4: *left panel: overlap of high-back vowels among JPN; right panel: overlap of high-back vowels among ENS.*

5.2. Vowel duration

The regression modeling arrived at a final minimal adequate model with a notable explanatory capacity accounting for 17.9 percent of the overall variability in the data. The data confirmed speaker type, vowel, and gender as significant main effects. More importantly, as this directly addresses the second research question and in addition to an interaction between gender and speaker type, the model reports significant interactions between speaker type and vowels. These latter interactions confirm that JPN exaggerate all vowel durations compared with the vowel durations of ENS.

Table 6: *Results of the mixed-effects regression modelling.*

	Final minimal adequate model		
Predictors	Est.	CI	p
(Intercept)	0.06	0.06 - 0.07	<0.001
type$_{JPN}$	0.03	0.02, 0.03	<0.001
label$_{i:}$	0.01	0.01, 0.02	<0.001
label$_{U}$	-0.00	-0.02, 0.01	0.587
label$_{U:}$	0.03	0.02, 0.04	<0.001
gender$_{male}$	0.01	0.00, 0.01	0.117
type$_{JPN}$ * label$_{i:}$	0.01	0.01, 0.02	0.002
type$_{JPN}$ * label$_{U}$	0.01	-0.01, 0.02	0.220
type$_{JPN}$ * label$_{U:}$	0.02	0.01, 0.03	<0.001
type$_{JPN}$ * gender$_{male}$	-0.02	-0.03, -0.01	0.003
Random effects			
ICC		0.08	
N		255 speakers 44 words	
Observations		3536	
Mar. R^2 / Cond. R^2		0.112 / 0.179	

The results show that JPN extend or exaggerate all high-front and back vowel durations and not just long vowels. The effect plot shown in Figure 5 furthermore shows that JPN exaggerate the duration difference of both /iː/ and /ɪ/ as well as /uː/ and /ʊ/. This is confirmed by Figure 6 as the differences between durations are notably higher for both vowel pairs among JPN compared with ENS.

The results of the statistical analysis thus confirm that JPN merge spectrally similar vowels and appear to compensate the lack of a differentiation by exaggerating both the durations of high-front and -back vowels and by exaggerating the durational

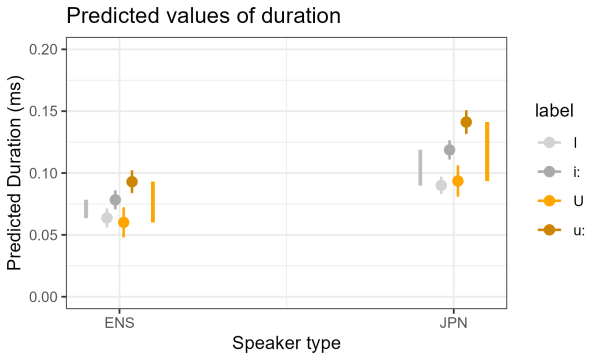


Figure 5: *Predicted duration values based on the final minimal adequate model by vowel and speaker type. (Gray and orange lines show the difference between short and long vowel durations within speaker groups)*

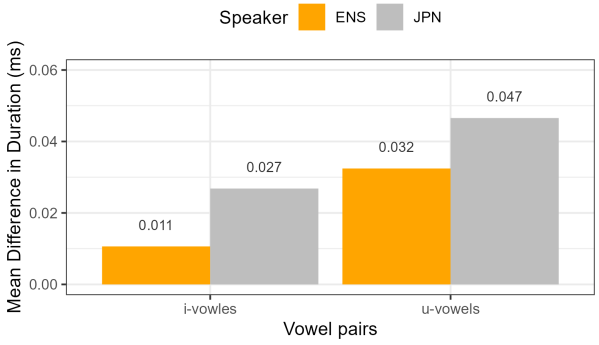


Figure 6: *Duration differences by speaker type (orange: ENS, gray: JPN) and vowel pair. The bars show the difference in duration between /iː/ and /ɪ/ (left) and between /uː/ and /ʊ/ (right).*

differences between long and short vowels (see Figure 6).

6. Discussion

The findings presented here confirm previous research which reported the tendency to merge spectrally close vowels produced by JPN in lab settings (see [20], [21]). Also in alignment with previous research [22] is that JPN exaggerated durational contrasts between spectrally similar vowels which, again, had been reported for JPN in lab settings.

The findings presented here also offer unique insights in that the study extends previous research to natural settings and substantially expands the empirical basis of existing research. In addition, the data analyzed here suggests a merger of high-back vowels (/uː/ and /ʊ/) among ENS in spontaneous speech.

One noteworthy limitation of the present study relates to the variable quality of the recordings which can be considered not only substandard but relatively poor for at least a subsection of the minute-long recordings that make up the spoken monologue component of the ICNALE. The reason for the low quality of the audio recordings is that the audio data were recorded predominantly using in-built microphones of mobile devices. While the quality of the audio data could, at least in part, be compensated using statistical procedures (kernel density estimation) which reduced the existing noise to a certain extent, such means are ultimately limited and unfit to fully remedy data quality. Another limitation consists in the fact that, given the variability and distributional characteristics of spontaneous speech, it is difficult to control the semantic and phonological environments of vowels, which, however, affect vowel production and thus formant values [23].

The advantages of the present study are that it has produced insights into vowel production by JPN in spontaneous speech which is under-explored even in learner corpus research. Also, the study is among the first to study JPN vowel production in natural settings which allows to generalize findings to real-life learner speech. Finally, and despite its limitations, the fact that the poor quality of the data could be compensated using advanced methods enables to extend the methods presented here to further automated corpus-based investigation on larger and more diverse samples of learner speech.

7. Conclusions

The present study represents one of the first large-scale, corpus-based studies of ESL vowel production in natural speech extending previous research by substantively extending the data base in terms of both the number of speakers and observations. In addition, the application of kernel density estimation to mitigate low quality audio data is promising but requires additional investigation and comparisons against gold standard data sets as well as manually annotated data to determine to what this method can compensate for poor data quality. Potential follow-up studies could zoom in on perception and intelligibility to investigate the auditory and cognitive implications of the acoustic effects presented here. Finally, the present study can be a prototype that can easily be extended to other learner varieties and multi-modal data sources.

8. References

[1] Gilakjani, A. P., and Ahmadi, M. R., "Why is pronunciation so difficult to learn?", English Language Teaching, 4(3), 74-83, 2011.

[2] Flege, J. E., "Second-language speech learning: theory, findings, and problems", in W. Strange [Ed.], Speech perception and linguistic experience: Issues in cross-linguistic research, 233-277, York Press, 1995.

[3] Franklin, A. D. and Stoel-Gammon, C., "Using multiple measures to document change in English vowels produced by Japanese, Korean, and Spanish speakers: The case for goodness and intelligibility", American Journal of Speech-Language Pathology, 23(4), 625-640, 2014.

[4] Homma, Y., "Acoustic phonetics in English and Japanese", Yamaguchi Shoten, 1992.

[5] Ladefoged, P., and Johnson, K., "A Course in Phonetics", Cengage, 2014.

[6] Ingram, J. C. L., and Park, S. G., "Cross-language vowel perception and production by Japanese and Korean learners of English", Journal of Phonetics, 25(3), 343-370, 1997.

[7] Kato, H., Tajima, K., Akahane-Yamada, R., "Native and non-native perception of phonemic length contrasts in Japanese", The Journal of the Acoustical Society of America, 110(5), 2686, 2001.

[8] Morrison, G. S., "Japanese listeners' use of duration cues in the identification of English high front vowels", in Larson, J. and Paster, M. [Eds.], Proceedings of the 28th annual meeting of the Berkeley Linguistics Society, 189–200, Berkeley Linguistics Society, 2002.

[9] Ishikawa, S., "Design of the ICNALE Spoken: A new database for multi-modal contrastive interlanguage analysis", Learner Corpus Studies in Asia and the World, 2, 63-76, 2014.

[10] Kisler, Th., Reichel, U. D. and Schiel, F. “Multilingual processing of speech via web services”, Computer Speech and Language 45: 326–347, 2017.

[11] R Core Team, “R: A language and environment for statistical computing”, R Foundation for Statistical Computing, Vienna, Austria. (https://www.R-project.org), 2022.

[12] RStudio Team, “RStudio: Integrated Development Environment for R”, RStudio PBC, Boston, MA (http://www.rstudio.com), 2022.

[13] Bořil, T., Skarnitzl, R., “Tools rPraat and mPraat”, in P. Sojka, A. Horák, I. Kopeček, and K. Pala [Eds.], Text, Speech, and Dialogue, 367-374, Springer International Publishing, 2016.

[14] Bombien, L., Winkelmann, R., and Scheffers, M., “wrassp: an R wrapper to the ASSP Library”, R package version 1.0.1, 2021.

[15] Wickham, H., et al., “Welcome to the tidyverse”, Journal of Open Source Software, 4, 43, 1686, 2019.

[16] Yang, B., “A comparative study of American English and Korean vowels produced by male and female speakers”, Journal of Phonetics, 24, 245–261, 1996.

[17] Deterding, D., “The Formants of monophthong vowels in standard southern British English pronunciation”, Journal of the International Phonetic Association, 27, 1-2, 47-55, 2009.

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

[19] Lüdecke, D., “sjPlot: Data visualization for statistics in social science”, R package version 2.8.10, 2021.

[20] Ueyama, M., “Duration and quality in the production of the vowel length contrast in L2 English and L2 Japanese”, in M. J. Solé., D. Recasens., and J. Romero. [Eds.], 15th International Congress of Phonetic Sciences, 1509-1512, Universitat Autònoma de Barcelona, 2003.

[21] Tsukada, K., “Native vs non-native production of English vowels in spontaneous speech: An acoustic phonetic study”, in P. Dalsgaard, B. Lindberg, and H. Benner [Eds.], Proceedings of the 7th European Conference on Speech Communication and Technology (Eurospeech 2001), 305-308, 2001.

[22] Tsukada, K., “Durational characteristics of English vowels produced by Japanese and Thai second language (L2) learners”, Australian Journal of Linguistics, 29(2), 287-299, 2009.

[23] Visceglia, T., Chiu-Yu, T., Kondo, M., Meng, H., and Sagisaka, Y., “Phonetic aspects of content design in AESOP (Asian English Speech cOrpus Project)”, Oriental COCOSDA International Conference on Speech Database and Assessments, 60-65, 2009.

Prosodic phrasing, pitch range, and word order variation in Murrinhpatha

Janet Fletcher[1,3], Evan Kidd[2,3], Hywel Stoakes[1,3], and Rachel Nordlinger[1,3]

[1]University of Melbourne
[2]Australian National University & Max Planck Institute for Psycholinguistics
[3]ARC Centre of Excellence for the Dynamics of Language
janetf@unimelb.edu.au;evan.kidd@mpi.nl;hstoakes@unimelb.edu.au;racheln@unimelb.edu.au

Abstract

Like many Indigenous Australian languages, Murrinhpatha has flexible word order with no apparent configurational syntax. We analyzed an experimental corpus of Murrinhpatha utterances for associations between different thematic role orders, intonational phrasing patterns and pitch downtrends. We found that initial constituents (Agents or Patients) tend to carry the highest pitch targets (HiF0), followed by patterns of downstep and declination. Sentence-final verbs always have lower Hif0 values than either initial or medial Agents or Patients. Thematic role order does not influence intonational patterns, with the results suggesting that Murrinhpatha has positional prosody, although final nominals can disrupt global pitch downtrends regardless of thematic role.
Index Terms: Murrinhpatha, intonation, pitch range, downstep, declination

1. Introduction

Murrinhpatha is a non-Pama-Nyungan polysynthetic language from the Daly River region of northern Australia (Figure 1). Like many other Indigenous Australian languages, it has flexible word order (see [1], and [2] for a general overview), with a recent psycholinguistic study confirming that Murrinhpatha has no default, underlying syntactic order for agent, patient and verb ([3]). In this paper we analyse phrasal intonation patterns produced by Murrinhpatha speakers from the same study to see whether this apparent lack of configurational syntax translates into prosodic flexibility in terms of intonational phrasing.

With perhaps the exception of various studies on Jaminjung (e.g. [4], [5]) the focus of previous work on intonation in Australian languages has tended to be on the interaction between prosody and information structure (e.g., [6], [7], [8]), with very little quantitative investigation of prosodic patterning associated with word order variation more generally. This study therefore has implications for our understanding of the interaction between post-lexical prosody and syntactic structure in Australian languages more broadly.

Whilst it is widely acknowledged that syntax has an influence on intonational phrasing, it is also generally accepted that it is mediated by prosodic structure [9, 10]. As shown in Australian languages with associated intonation system analyses, the intonational phrase boundaries and clause boundaries typically align (see [11] for a general overview). However, some important exceptions have been observed in languages like Wubuy [12] and Dalabon [13], where there can be a mismatch between prosodic phrasing and grammatical word hood. Other languages can show a high level of prosodic integration of complex verbal elements, particularly in highly agglutinative polysynthetic languages such as Bininj Kunwok [14].

Figure 1 *Map of Northern Australia showing Wadeye where Murrinhpatha is predominantly spoken.*

Various features of Murrinhpatha phrasal prosody have been described previously in [15]. Nominals are usually realised in separate prosodic phrases. Verbal elements that combine with subject or object elements show flexible order and can be realised as multiple prosodic phrases i.e. there is a degree of prosodic and syntactic independence. This fits well with the core finding reported in [3], that there is no underlying default order for thematic roles in Murrinhpatha. It also suggests that regardless of participant order, we will observe the normal varied patterns of prosodic or intonational phrasing that are found in studies of narrative discourse for other Australian languages [4, 14, 16], with no particular pattern associated with either Agent- or Patient-initial or -final order. It is further predicted that pitch downtrends and phrasal contours will reflect commonly observed patterns in other Australian languages. For example, it has been shown that, regardless of word order, there is default "positional" prosody whereby the first constituent bears the highest f0 target in a sequence of intonational phrases [11, 16], and subsequent normal downtrends (i.e. pitch declination) across the rest of the utterance are observed. If positional prosody is the norm, then we would not expect to see any major deviation from this pattern among the different thematic orders. On the other hand, previous investigations of word order variation in Australian languages have revealed evidence of local pitch range reset in final nominals when they are realised as separate intonational

phrases (e.g. [13], [16]). It is therefore also important to consider intonational phrasing patterns and their interaction with pitch range trends. In general, the intonational typology of the language appears to be phrasal with strongly delimitative edge-marking pitch patterns.

2. Method and Materials

2.1. Participants

The corpus was recorded in Wadeye in the Northern Territory across three fieldtrips in 2016 and 2018 as part of a previous sentence production experiment [3]. The dataset described here is a subset of the original corpus and includes 637 utterances from eighteen adult speakers of Murrinhpatha (8 males, 10 females). Murrinhpatha is the first language for all participants.

2.2. Materials and Procedures

Participants were asked to describe 48 pictures of various transitive events that were presented on a laptop, interspersed amongst 96 fillers (mostly depicting intransitive events). Agent and patient humanness (+/- human) was manipulated in a 2x2 design. Two examples of pictures used in the original study are shown in Figure 2.

Figure 2 *Example of two pictures used in the naming task*

Pictures were randomized and presented to participants across four blocks. Each picture was mirror-reversed such that the agent appeared to the left for half of the participants and to the right for the other half. A portable eye-tracker attached at the bottom of the screen captured their eye-movements (for more details see [3]), and participants' verbal descriptions of the actions in each picture were recorded using a Zoom H4n recorder.

2.3. Data processing and analysis

The data analyzed in this paper were chosen based on the quality of the original recordings. The recorded speech files were transcribed and glossed orthographically in ELAN and coded for thematic word order e.g. Agent – Verb – Patient (AVP) or Patient – Verb – Agent (PVA) by the third author. Annotated files were exported from ELAN into Praat, and word and phoneme-level segmentation was performed using the Montreal Forced Aligner [17]. A hierarchical database was constructed using the EMU speech database management system [18] using tiers for utterance, word, thematic role, phoneme, and intonational tone target. Each utterance was annotated manually using an adaptation of a simple Autosegmental-Metrical model of intonation, where peaks in the f0 contour were labelled H* and obvious troughs or terminal points of falling tunes at the right edge of a word or group of words was labelled L% [19] denoting an Intonational Phrase (IP) boundary.

Obvious downtrends were captured using the ! diacritic to represent downstepped !H* pitch accents. The majority of right-edge demarcative pitch movements at intonational phrase (IP) boundaries in the dataset were falling (H* L%), in keeping with earlier observations of Murrinhpatha phrasal prosody [15]. Other right boundary contours that were mid to high level (H* H%) were also sometimes observed in utterance-medial positions. There were also many cases of minor intonational phrases (iP) that had either falling (H* Lp) or level tunes (H* Hp). Minor intonational phrases show a smaller degree of juncture than major IPs but observable pre-boundary lengthening and a pattern of downstep within the major IP. Pauses were also annotated. These only occurred between IPs.

An example of an annotated f0 contour from the corpus is shown in Figure 3. The utterance shows an example of Agent-Verb-Patient thematic role order and is produced in two major intonational phrases separated by a pause. The first intonational phrase (IP) illustrates a clear downtrend from the Agent to the verb complex, which also shows localized downstep between the first and second pitch accents. There is a pause after the verb complex and an upwards pitch range reset for the final intonational phrase that includes the Patient.

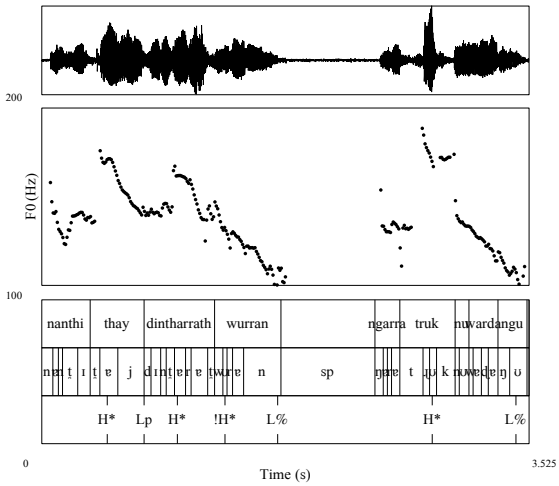


Figure 3 *F0 contour showing downstep in the first intonational phrase produced by a male Murrinhpatha speaker for an AVP order utterance:*

Nanthi thay dintharrath=wurran ‖ ngarra truk-nu-warda-ngu
CLF:THING tree 3sgS.nFut.fall on=3sgS.nFut.go LOC truck-DAT-NOW-DM
"The tree is falling onto the truck"

F0 values were extracted for the highest f0 targets associated with pitch accents within each IP or iP akin to a ToBI-style HiF0 measure. They were converted to semitones using a 50 Hz benchmark and were normalized as z-scores to enable comparison across male and female speakers. Patterns of intonational phrasing were also noted for each utterance in relation for different thematic orders, as well as patterns of pausing. The normalized f0 values were included in a mixed effects model with fixed factors: Part of Speech (agent, patient, verb) and word order pattern with random effects of speaker and word, using *lmertest* and *step* in R [20]. Post-hoc Tukey tests using the R package *emmeans* were used to investigate any interactions more closely.

3. Results

3.1. Thematic role distribution

Table 1 summarizes the distribution of thematic role orders in the subset of data analyzed in this paper. In keeping with the results reported in [3] for the full corpus, Agent-initial word orders were the most common, followed by Patient-initial word orders. The full corpus contained all possible combinations of word orders except for VPA [3].

Table 1. *Word order distribution.*

Word Order	Number
AVP	306
APV	123
PVA	61
PAV	54
AV	33
PV	49
VAP	9
Total	**637**

3.2. Pitch range

Figures 4 and 5 plot the highest normalized f0 target values associated with H* accents (i.e. HiF0) across each thematic component for the dominant utterance types: Agent – Verb – Patient (AVP) and Patient – Verb – Agent (PVA). For both thematic role orders, it is clear that the initial element, regardless of whether it is an Agent or Patient, contains the pitch accent with the highest f0 value (HiF0) for the utterance. There is a statistically significant interaction between part of speech and participant order (F=9.1423; p<0.0001).

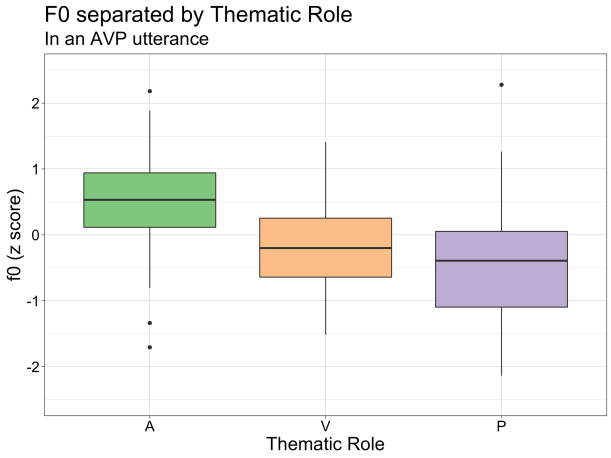


Figure 4 *Normalized f0 values for the highest f0 targets (H*) in Agent-Verb-Patient ordered utterances*

In general, utterance-final elements carry lower H* pitch accents than words that are either an initial Agent or Patient. However, the most significant downtrend is from the initial Noun (Agent) to the following Verb (AVP t=5.378). This is where most instances of downstep are observed in this frequent word order. The main driver of the interaction between word order and thematic role appears to be driven by the location of the Verb: Verb-final utterances have lower H* pitch accents regardless of whether preceded by an Agent and/or Patient. In fact, the magnitude of utterance-level (i.e., global) downtrend tends to be greater in Verb-final utterances, as shown in Figure 6 for APV sequences, and a similar pattern is observed in PAV word order (APV t=5.34 p<0.0001; PAV t=5.318; p<0.0001). In PV and AV word orders, there are also trends in the same direction (e.g. PV t=3.54 p<0.07), but the differences are not statistically significant. This is probably because speakers often produce these utterances with a lot of intervening or following material, given the free nature of the speech task.

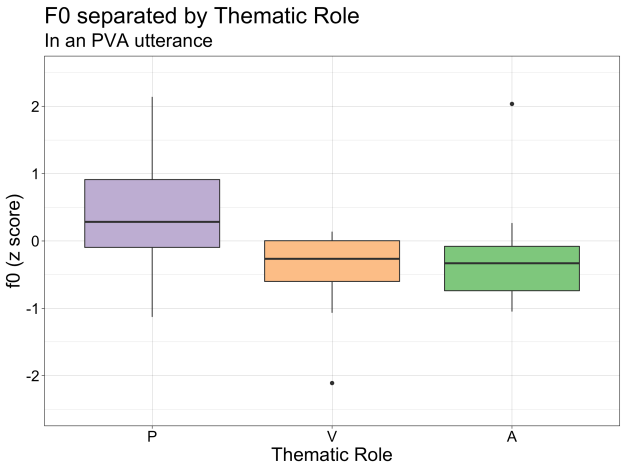


Figure 5 *Normalized f0 values for the highest f0 targets (H*) in Patient-Verb-Agent ordered utterances.*

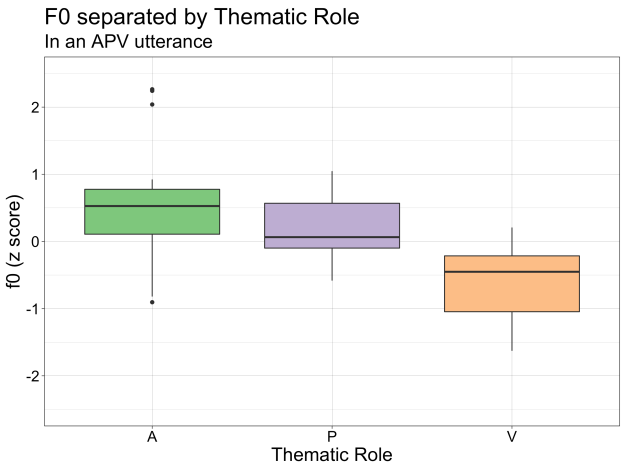


Figure 6. *Normalized f0 values for the highest f0 targets (H*) in Agent-Patient-Verb ordered utterances*

3.3 Intonational Phrasing

A complex picture emerges when intonational phrasing patterns are examined. Given the relatively free nature of the task, a range of different intonational phrasing patterns are observed as one would expect. Table 2 summarizes some key observations for a portion of the dataset for those utterances that have both Agents and/or Patients. Column 1 documents the proportion of

utterances where the final A, V, or P is realized as a separate major IP with final pitch reset, whereas column 2 shows the proportion of cases where the first two constituents are grouped into a major IP.

Table 2. *Intonational phrasing distribution and Thematic Role order for utterances including Agents and/or Patients*

Order	*Final Pitch reset of A,P, or V – separate IP (%)*	*Grouping of first two constituents (%) into IP*
PVA	40	75
AVP	66	77
APV	44	47
PAV	25	35
VAP	25	25
AV	7	26
PV	17	50

Verb-medial utterances (i.e., AVP or PVA) group the Verb with the initial Agent or initial Patient into a major intonational phrase (IP) in 77% and 75% of cases, respectively. In the remaining 23% or 25% of utterances, the A, V or P elements are realized as either separate IPs, or iPs within a single IP depending on the speaker. For the cases where either the A or P groups with the following V, the A and V or P and V are also sometimes realized as separate minor iPs within the major IP. Pre-boundary lengthening is usually observed in the IP-final syllable that marks the right edge of the Agent or Patient +Verb complex and there is also an optional pause in many cases (29% for AVP, 20% for PVA utterances). A local downtrend is often observed within the initial IP with Verbs that are part of the same major IP, typically bearing a downstepped !H* pitch target or showing a pitch downtrend (as reflected in the example in Figure 3, and shown in the overall f0 target plots in Figures 4 and 5). Thus, there appears to be a degree of prosodic cohesion, as evidenced by the downtrend patterns within the A+V or P+V IPs *regardless of thematic role* for AVP and PVA utterances.

In addition, the final element in verb-medial utterances (either a P or A) is often realized as a separate major intonational phrase in 40% and 66% of cases respectively (Table 2, column 1). This occurs regardless of whether there has been initial grouping of the AV or PV as discussed above. In these cases, the pitch topline (HiF0) of the final IP is marginally lower or at the *same level* as the HiF0 of the initial major intonational phrase of the utterance, i.e. where the typical global declination pattern is disrupted. An example of this is the contour illustrated in Figure 3, where the speaker has emphasized the final noun and produced the utterance-final Patient in an expanded pitch range that results in a HiF0 (H*) value that is equivalent to the HiF0 value on the initial Agent. The rest of the cases analyzed in this subset of data follow predictable global downtrend patterns, i.e., there is declination across the entire utterance.

For less frequent utterance types that include both thematic roles, more variation is observed with a range of different phrasing patterns produced. Interestingly, in 47% and 35% of PAV and APV utterances, final verbs are grouped together with the preceding thematic role (Agent or Patient) into a single intonational phrase, with the A or P realized in a separate initial IP. Pre-boundary lengthening of IP final syllables and pausing between major phrases are observed in many cases but the A and P often have a similar HiF0 value and we do not observe a strong pitch downtrend across the two. By contrast, final Verbs are often realized with a significantly lower HiF0 than preceding constituents (as shown in Figure 6 and as confirmed statistically). In the relatively rare VAP pattern (9 in this subsection of the original corpus), speakers tend to produce all three constituents as three separate major or minor intonational phrases. This is also more likely to be the case in PAV utterances. For utterance-types that consist of an Agent-Verb (AV) or Patient-Verb (PV) sequence, there are also a variety of intonational patterns as speakers often inserted a degree of extraneous material in their utterances, giving rise to multiple phrasing patterns. In PV and AV utterances, for example, 26% and 50% are produced as a single IP (Table 2, column 2) with the rest produced as separate IPs. In the shorter PV and AV utterances final verbs are usually produced in a lower overall pitch range similar to other verb-final word orders.

4. Discussion

The Murrinhpatha speakers analysed in this preliminary study produce a variety of intonational phrasing patterns regardless of thematic word order, showing that in the same way as there is no underlying default word order in Murrinhpatha [3], there is no predictable intonational phrasing pattern that reflects one thematic order rather than other. Our results also support the claim that syntax and prosody are largely independent in Murrinhpatha [15]. Speakers produce a wide range of intonational phrasing patterns for each participant order type, although the most common thematic orders analysed in this study also show a similar intonational profile in that initial agents or patients are grouped with following verb complexes into major intonational phrases with final patients or agents realised as separate IPs.

Pitch downtrends are largely similar across the corpus regardless of thematic role order. Speakers tend to expand their pitch range for the initial intonational phrase of the utterance regardless of whether the utterance begins with an Agent or Patient. Notably, the lowest HiF0 values are typically registered on Verbs across the dataset, and are clearest when Verbs are utterance-final. Intonational phrases tend to show typical downtrends, particularly in APV word order with final Verbs realised with lower HiF0 targets than final Patients in AVP order, for example. Conversely, in AVP and PVA utterances, there are clear downtrends between the Agent or Patient to the Verb, but these are often disrupted when there is a following Patient or Agent. It should be noted, however, that all types of Verb-initial utterances analysed in [3] (i.e. VA or VP orders) were not included in the intonational analysis reported here and should be analysed in any future work.

This quantitative investigation of Murrinhpatha confirms that the language has phrasal prosody with demarcative right-edge pitch movements that are largely falling, as suggested in [15], although we also observe mid-level tunes in utterance medial contexts. Like many other Australian languages whose intonation systems have been quantitatively analyzed so far, Murrinhpatha prosody is indeed "positional" in that utterance-initial IPs typically utilise the widest pitch range of an utterance. Speakers can elect to manipulate both phrasing and pitch range downtrends when there are final nominals, but this appears to be largely independent of participant role.

5. References

[1] Hale, K. “Warlpiri and the grammar of nonconfigurational languages”. Natural Language and Linguistic Theory 1:5-47.1983.

[2] Nordlinger, R. “Constituency and grammatical relations in Australian languages.” In H. Koch & R. Nordlinger (Eds.) The Languages and Linguistics of Australia: A comprehensive guide. Berlin: De Gruyter. 215-262. 2014.

[3] Nordlinger, R., Rodriguez, G.G., & Kidd, E. “Sentence planning and production in Murrinhpatha, an Australian 'free word order' language”. Language 98(2), 187-220. 2022.

[4] Simard, Candide. “The prosodic contours of Jaminjung.”. Pacific Linguistics, in press.

[5] Schultze-Berndt and Simard C. “Constraints on noun phrase discontinuity in an Australian language: the role of prosody and information structure”. Linguistics 50, 1015-1058. 2012.

[6] Simpson. J. and Mushin I. “Clause-initial position in four Australian languages”. In Mushin, Ilana & Brett Baker (eds.) Discourse and grammar in Australian languages. Amsterdam: John Benjamins. 25-57, 2008.

[7] Fletcher, J., Stoakes, H., Loakes. D., Singer, R. “Intonational correlates of subject and object realisation in Mawng.” Speech Prosody 2016.188–192. 2016.

[8] Mushin, I. “Word order pragmatics and narrative functions in Garrwa”. Australian Journal of Linguistics, 25:2, 253-273.

[9] Nespor, M. and Vogel, I. “Prosodic phonology”. Dordrecht: Foris, 1986.

[10] Selkirk, E. (2011). “The Syntax-Phonology Interface.” In J. Goldsmith, J. Riggle, and A. Yu, eds., The Handbook of Phonological Theory, 2nd edition, 435-484. Oxford: Blackwell.

[11] Fletcher J, & Butcher, A. “Sound patterns of Australian languages. In Koch, H. & Nordlinger, R. The Languages and Linguistics of Australia: A comprehensive guide. Berlin: De Gruyter.89-132. 2014.

[12] Bundgaard-Nielsen, R. and Baker, B. “Pause acceptability indicates word-internal structure in Wubuy”. Cognition, 19, 104167. 2020.

[13] Evans, N., Ross, B., Fletcher J. “Big words, small phrases: mismatches between pause units and the polysynthetic word in Dalabon”. Linguistics 46-1, 87-127. 2008.

[14] Bishop, Judith. “Aspects of intonation and prosody in Bininj Gun-wok: An autosedmental-metrical analysis”. (Doctor of Philosophy Unpublished PhD Thesis), University of Melbourne, 2002.

[15] Mansfield, J. “Murrinhpatha morphology and phonology”. Berlin:de Gruyter (Pacific Linguistics), 2019.

[16] Jepson, Katie. “Prosody, prominence, and segments in Djambarrpuyŋu”. Unpublished PhD Dissertation University of Melbourne. 2019.

[17] McAuliffe, Michael, Michaela Socolof, Sarah Mihuc, Michael Wagner, and Morgan Sonderegger. “Montreal Forced Aligner: trainable text-speech alignment using Kaldi.” Proceedings of the 18th Conference of the International Speech Communication Association. 2017.

[18] Winkelmann, R., Jänsch, K., Cassidy, S., Harrington, J. “emuR: Main package of the EMU Speech Database Management System.” R package, Version 1.0.0. 2018.

[19] Ladd, D.R. “Intonational Phonology”. Cambridge: Cambridge University Press. 2008.

[20] 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.

A preliminary study of lexical pitch accents in the Split dialect of Croatian

Marija Tabain[1], Mate Kapović[2], Matthew Gordon[3], Adele Gregory[1], Richard Beare[4]

[1] La Trobe University, [2] University of Zagreb, [3] University of California Santa Barbara, [4] Monash University and Murdoch Children's Research Institute

m.tabain@latrobe.edu.au, mkapovic@ffzg.hr, mgordon@linguistics.ucsb.edu, a.gregory@latrobe.edu.au, richard.beare@monash.edu

Abstract

The pitch accent system of the Split variety of Croatian exhibits an intermediate stage of the Neo-Štokavian retraction of stress, with the so-called "double accent" as the result. In this study, we examine acoustic data from eight female speakers of this variety. Although duration data point to stress on the initial syllable of the word – in line with the modern-day standard – energy data are more ambiguous. In addition, the rising lexical pitch accent is found to have a significant fall on the post-tonic syllable – a situation which suggests the separation of pitch from durational cues to stress.

Index Terms: lexical pitch accents, Croatian, dialectology, tonogenesis, sound change

1. Introduction

Croatian is a South Slavic language that, in the standard variety, is similar to the other officially recognized national languages Bosnian, Serbian and Montenegrin (collectively known as BCMS). Other South Slavic languages include Slovene, Macedonian and Bulgarian. These South Slavic languages, together with the Western and Eastern Slavic languages, are in turn descended from Proto-Slavic [1, p. 15].

In this paper, we consider the Split variety of Croatian [2], [3 p. 470-473]. Split is the second largest town in Croatia, with more than 160.000 inhabitants. In particular, we focus on the lexical pitch accents (LPAs) of the language, which in Croatian involve tonal alignment with the stressed syllable of the prosodic word.

Historically, BCMS is well-known for the leftwards retraction of stress, with the end result that the vast majority of words in the modern lexicon have stress (largely marked by extra duration) on the first syllable of the word [4, p. 671-733]. Consider the female name 'Maria' (in Croatian *Marija*), which in other European languages has stress on the second syllable:

ma.RI. ja

However, in Croatian, this word is stressed on the first syllable:

MA. ri. ja

Notably, this retraction of stress led to the creation of a contrast between rising and falling LPAs in the so-called Neo-Štokavian dialects, which are the basis for the standard language (in this paper, we also refer to the LPAs as "prosodemes"). Since historically a High tone was located on the stressed syllable of the word, the retraction of stress without a concomitant retraction of the High tone resulted in a pitch contrast that had not previously existed. In the case of 'Maria', it led to a rising pitch accent on the word (marked traditionally as *Màrija* in standard Croatian – note that the Croatian accentual notation does not correspond to the IPA), since the high tone remained aligned with the second syllable /ri/, while the stress moved to the initial syllable /ma/.

MA. ri. ja
L H

In modern BCMS accentology, a Low tone is assigned to the first syllable of the rising prosodeme, to indicate a distinct low target on the stressed syllable – in a ToBI style notation [5], this has been denoted as L*+H (however for a different interpretation, see e.g. [6], [7]).

By contrast, a word such as *jȁbuka* 'apple' has a falling pitch accent, whilst also having stress on the first syllable:

JA. bu. ka
H L

In this case, the High tone is aligned on the first syllable, but it also combines with a following Low tone to ensure that a fall occurs on the stressed syllable. In Godjevac's ToBI analysis, this is denoted as H*+L.

Significantly, in the Split dialect of Croatian, the rising prosodeme has a variant that is referred to as a "double accent". In this case, there is a perception of a rising accent on the initial, stressed syllable, followed by a fall on the post-tonic syllable. In the case of *Màr̀ija* (with both a rising accent on the first syllable and a falling accent on the second syllable), one could hypothesise the following tonal associations:

MA. ri. ja
L H L

This "double accent" variant is considered a remnant of the historical state of the language before the leftwards retraction of stress [8, p. 9–11], [3, p. 470–471], [4 p. 49 (f. 107), p. 673, p. 686]. The old dialect of Split (one of the so-called Čakavian varieties) was characterized by older stress placement, closer to Proto-Slavic. The modern Split dialect is no longer Čakavian, but instead heavily Neo-Štokavianized due to the influence of the Neo-Štokavian hinterland and the standard dialect (Neo-Štokavian dialects have full stress retraction). However, the older Čakavian variants such as *Mar̀ija* (with a falling accent on the second syllable) can still be heard from older speakers in Split, alternating with the double accent. Younger speakers alternate between the standard *Màrija* and the double accent.

In this study, we explore the duration, energy and pitch properties of the Split dialect prosodemes, with the aim of trying to understand the historical stages between stress retraction and the creation of distinct lexical prosodemes.

1.1. Some additional information on LPAs in Croatian

In addition to the contrast between rising and falling pitch accents, standard Croatian has a vowel length contrast which in traditional accentology combines with the pitch accents to create a 4-way contrast: Short Rising (SR), Short Falling (SF), Long Rising (LR) and Long Falling (LF). Although Croatian orthography provides accentual markings for all four prosodemes, in practice they are not written except to avoid

ambiguity, and most readers are not familiar with the accentual system. It should be noted that the vowel length contrast in principle carries a high functional load in both standard Croatian and in the Split variety, both lexically and grammatically (e.g. nominative versus genitive case). The system of pitch accents is likewise greatly affected by grammatical elements such as noun case. However, there are very few minimal pairs involving pitch only, and even fewer that involve the same grammatical class (using the traditional accentual notation, examples of nouns in the nominative include *pàra* (SF) 'steam' : *pàra* (SR) 'money', *Lûka* (LF) 'Luke' : *lúka* (LR) 'port').

It is also important to note that the contrast between rising and falling accents exists almost exclusively on polysyllabic words with stress on the first syllable. This is because monosyllabic words almost always have a falling accent (given that the pitch peak in the rising accent occurs on the second syllable – exceptions involve the deletion of a final vowel); and polysyllabic words with stress on a non-initial syllable almost always have a rising accent (in line with the historical situation for the language) – e.g. *iznòsiti* 'to carry out', which has a Short Rising pitch accent and stress on the second syllable. For this reason, in the present study, we only consider polysyllabic words with stress on the initial syllable.

Figure 1: *Lexical pitch accents of standard Croatian.*

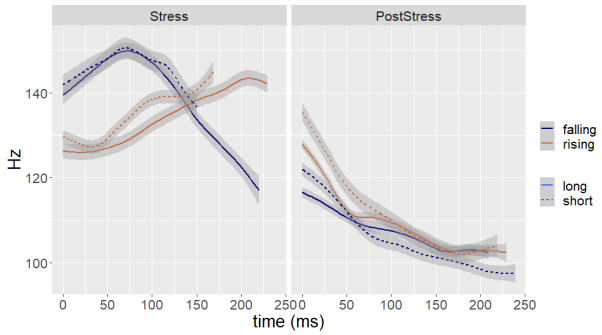


Figure 1 shows a typical pattern for the lexical pitch accents of standard Croatian. The rising and falling pitch accents are denoted by colour differences, and the length contrast is denoted by linetype. These data are based on 256 word tokens produced by the second author of this paper, a speaker of Standard Croatian, using the same set of stimuli as were used for the recordings of Split speakers (details below).

It can be seen that the falling prosodemes have a clear fall on the initial stressed syllable. The rising pitch accents have a rise over the course of the initial stressed syllable; however, the starting point of the rise is higher than the end point of the falling pitch accent, due to the presence of a low phrasal tone at the right-edge of the single-word utterance for this speaker. In the post-stress syllable, the rising prosodeme begins slightly higher than the falling prosodeme, but within 50 ms of the post-stress vowel, the differences are almost non-existent (these data have been GAM-smoothed for ease of interpretation).

Figure 2 shows boxplots for duration of the stressed vowel (left panel) and the post-stress vowel (right panel), for the same data as was shown in Figure 1. It can be seen that while the short vowels in the stressed syllable tend between 100 and 150 ms, the long vowels tend between 150 and 200 ms. Importantly, there is a tendency for the rising prosodemes within each length pair to be longer than the falling prosodemes. It is usually assumed that this is due to the greater time required to achieve a rise in pitch, as opposed to a fall in pitch [9]. By contrast, vowel duration in the post-stress syllable tends to be between 50 and 100 ms (the great variability in this panel is due to the fact that the data only represent one speaker, and we did not control for post-tonic phonemic vowel length).

Figure 2: *Duration of stressed and post-stressed vowels in standard Croatian.*

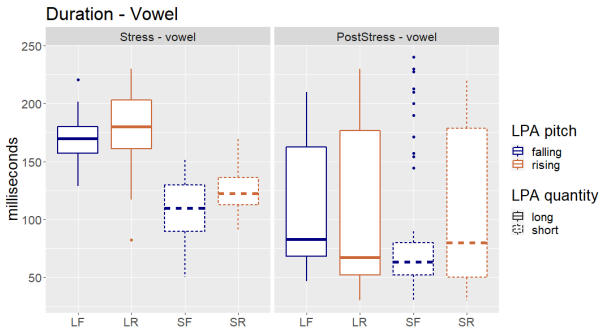


Figure 3: *Mean RMS Energy of stressed and post-stressed vowels in standard Croatian.*

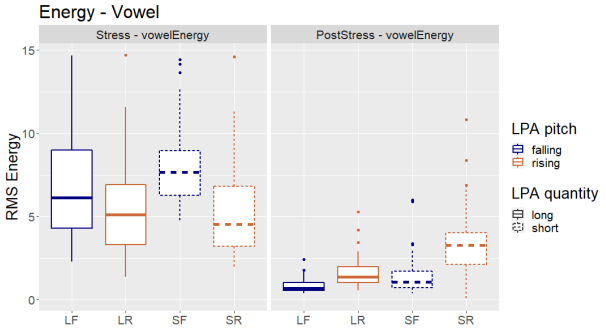


Figure 3 shows the mean RMS energy calculated across the entire vowel for this same set of data. It can be seen that the stressed vowel (left panel) clearly contains more energy than the post-stressed vowel (right panel), thus providing further evidence of lexical stress on the initial syllable of the word. It may also be noted that the falling pitch accents have more energy than the rising accents for the stressed vowel, while the opposite pattern holds for the post-stressed vowel. This mirrors the overall pitch levels across the respective syllables.

In the remainder of this paper, we present the prosodemes of the Split dialect, and consider to what extent they reflect the historical situation of stress retraction.

2. Method

2.1. Speakers and recordings

Eight female speakers of the Split variety were recorded in March 2022 at the Department of Phonetics recording studio at the Faculty of Humanities and Social Sciences, University of Zagreb, under the supervision of a recording technician and the second author. All speakers were born in Split between 1996 and 2002 (with most born between 1999 and 2002), and had arrived in Zagreb for their university studies between 2015 and 2021.

It should be noted that although the Split variety is not the standard, it is in fact a prestige and well-liked variety in Croatia, associated with popular music and sporting personalities. Thus, the speakers in our recordings were not using a stigmatized variety of the language in the university setting, and they were encouraged to pronounce the words as they would normally pronounce them. The second author of the study verified that the speakers spoke the Split variety of Croatian during the recording.

Speakers were recorded using an AKG C414 B-ULS microphone and RME Fireface UFX soundcard. Recordings were saved as mono WAV files using a 44.1 kHz sampling rate and 16 bit bit-depth.

2.2. Stimuli

Speakers read a list of 65 real Croatian words that illustrated the four prosodemes. Words were read in isolation, with no carrier phrase. The words were either stand-alone lexical items, or prosodic words with enclitics (BCMS has an extensive system of enclitics and proclitics). Words contained at least three syllables, and up to six syllables (17 words with three syllables, 24 words with four syllables, 20 words with five syllables and four words with six syllables). The purpose of having longer words was to make sure that there were at least two syllables following the initial stressed syllable, in order to allow for any phrasal tones that may occur in these single-word utterances.

The list contained:

11 Long Falling
17 Long Rising
17 Short Falling
20 Short Rising

The number of tokens in each prosodic category is unbalanced, since for the purposes of a related study, we tried to find words that had different syllable structures for each prosodeme (i.e. open ~ closed by a sonorant ~ closed by an obstruent). We based these decisions regarding syllable structure on the Maximum Onset Principle, and on a word-game that allows for the reversal of syllables. It should be noted, however, that most of the words have an open syllable structure on the stressed syllable, given that BCMS is highly permissive in terms of syllable/word onset clusters.

The wordlist was randomized and each speaker read four repetitions of the list. One speaker did not produce the word *klȁckali ʃu ʃe* (Short Falling) 'they were see-sawing', due to the addition of this word to the list after the speaker had been recorded. This gives a total of 2076 word tokens for analysis (65 x 4 repetitions x 8 speakers, minus four tokens). Most of the words contained a vowel as the nucleus, but four words contained a syllabic /r/ as the nucleus (this is a schwa-like vowel with usually just one very brief period of closure).

One final note concerns vowel length on the post-tonic syllable. As mentioned above, BCMS contrasts long and short vowels in all positions of the word (except before the stressed syllable), and the question of post-tonic length is important in both a synchronic analysis of dialects, and in terms of historical distinctions. In the present study, we were not able to control for post-tonic length of the vowel, given the other factors we attempted to control for. For this reason both long and short vowels occur in the syllable following the stressed syllable in our word-list.

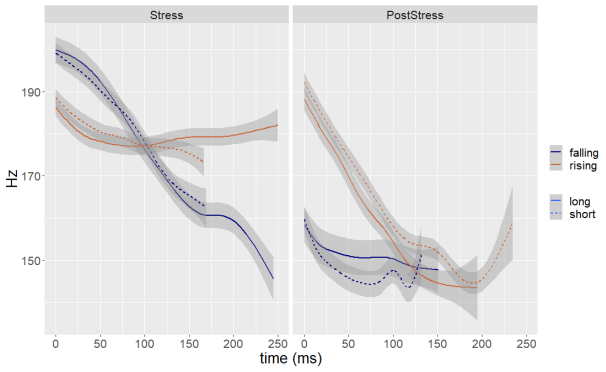
2.3. Data analysis

Phonetic transcriptions of the words were imported from a spreadsheet and used for preliminary phonetic segmentation with the Munich AUtomatic Segmentation system (MAUS – [10]) pipeline function G2P->MAUS->PHO2SYL. Manual correction of the phonetic MAUS labelling was conducted using the EMU Speech Database Management System [11, 12], interfaced with the R statistical software package [13]. The Snack signal processor [14] was used for calculating formants. The VoiceSauce software was used to extract voicing measures including Straight f0 for the pitch traces shown below [15], as well as RMS energy. Plots were generated using the *ggplot2* package in R [16].

3. Results

Figure 4 shows the lexical pitch accents for the Split variety – this figure is parallel to Figure 1, but shows data for eight speakers instead of only one speaker (note that no speaker normalization has been carried out on these data). It can be seen that both the rising and falling prosodemes have a relatively high f0 at the left edge of the stressed vowel. The falling prosodemes (blue lines) start very high, and fall very quickly over the course of the stressed vowel, with the fall completed before the start of the post-stress vowel. By contrast, the rising prosodemes (brown lines) also start high, but then fall slightly; the long rising prosodeme then starts to rise during this stressed vowel. This is followed by a high fall on the post-stress vowel for this rising prosodeme. This high fall is perhaps indicative of the "double accent" that is perceived on many instances of the rising prosodeme in the Split variety (i.e. the one perceived to have a following fall).

Figure 4: *Lexical pitch accents for Split variety.*



It is important to note that in this plot, we have not separated the perceived "double accent" rising variant from the perceived regular rising variant. Despite this limitation, it is clear that the Split pitch traces are quite different from the pitch accent traces presented in Figure 1. There is evidence of a high tone at the left edge of the stressed vowel in the Split variety. This high tone is most likely postlexical, but as a result, some researchers hear the rising accent as more "level" in this variety. It also results in the falling prosodeme (blue lines) having an extra high starting point. By contrast, in the variety shown in Figure 1, this falling prosodeme has an initial rise, before a fall on the stressed syllable (completed before the start of the post-stress vowel) - cf. e.g. Škarić in [17, p. 321–322].

As already mentioned, the rising prosodeme (brown lines) does not have a very distinct rise over the course of the stressed vowel in the Split variety, partly due to the presence of the post-

lexical high at the left edge. By contrast, the variety shown in Figure 1 has a much more distinct rise over the course of the stressed vowel, due to a lower initial starting point.

Please note that the great variability seen in the pitch contours at the right edge of the post-stressed vowel, especially the short vowels (dotted lines), is likely due to the different intonation patterns adopted by different speakers – some adopt a final fall, some a rising contour, and some a level contour.

Figure 5: *Duration of stressed and post-stressed vowels in Split variety.*

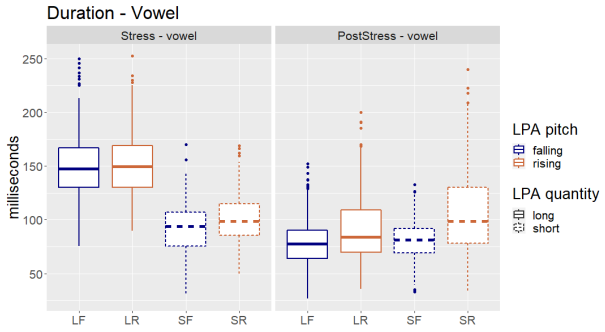


Figure 6: *Mean RMS Energy of stressed and post-stressed vowels in Split variety.*

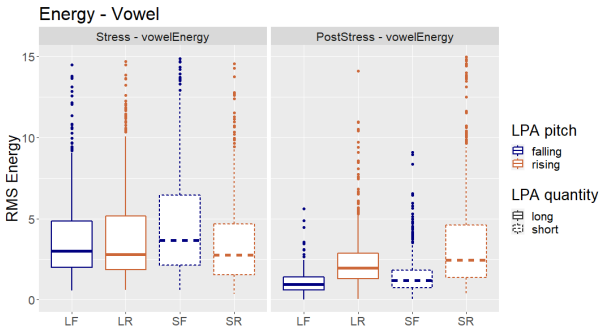


Figure 5 shows the duration data for stressed and post-stress vowels for the Split variety – this figure is parallel to Figure 2. Similar durational differences between the long and short prosodemes are once again evident in the stressed vowel (left panel). However, it seems that the differences in duration between rising and falling prosodemes in the stressed vowel are not so obvious in the Split variety as they are in some other varieties. It would also appear that there is not quite so much difference between stress and post-stress vowel duration in the Split variety (at least, comparing short stressed vowels with post-stressed vowels). However, given that we have not been able to control for post-tonic vowel duration, this observation is very tentative; it should also be noted that post-tonic vowel length mostly occurs following rising prosodemes (in both the standard and in the Split variety of Croatian).

Figure 6 shows the mean RMS energy for stressed and post-stress vowels for the Split variety – this figure is parallel to Figure 3. It can be seen that the difference in energy between the stressed vowel and the post-stress vowel is not as great in the Split variety as in the standard. Indeed, a calculation of differences in energy between the stressed and post-stress vowel gives a mean value of 4.2 dB for the variety deemed the standard here, but only 1.2 dB for the Split variety. This suggests that one of the standard cues to lexical stress is not as strong in the Split variety, a fact which may contribute to the perception that it is more difficult to locate the "stressed" syllable in this variety. It may also be noted that whilst the rising prosodemes have more energy than the falling prosodemes in the post-stress syllable in the Split variety, any differences are less clear in the stressed syllable. This difference in the post-stress syllable is possibly a reflection of the greater differences in pitch in the post-stress syllable as compared to the stressed syllable.

4. Discussion

Our results confirm the possibility of a "double accent" for the Split rising prosodemes (brown lines). There is indeed a high fall on the post-tonic syllable (as opposed to a mid fall in the standard variety), following an initial rise on the stressed syllable. Notably, this initial rise is not quite as extensive as the rise that is typically observed for the standard variety (as exemplified by the second author of this paper).

A second finding of the present study concerns a possible post-lexical High tone at the left edge of the prosodic unit (word or phrase) in the Split variety. This possible High tone serves the function of raising the overall pitch on the initial syllable. As a result, the falling prosodeme (blue lines) begins extra high, without the initial slight rise in pitch that is seen in the standard variety. It is possible that this initial High interacts with the lexical pitch accents, either adjusting their timing, or adjusting the f0 target. To what extent such an initial post-lexical High tone may play a role in diachronic changes in stress and lexical pitch accents is an interesting consideration.

A third finding of our study concerns the important cues to lexical stress, namely duration and energy. Duration results confirm that lexical stress is located on the initial syllable of the words in the Split variety, as is also the case for the standard Croatian language. However, the energy results for Split did not show quite the same difference between the stressed and post-stressed syllables that were evident in the "standard" speech of the second author. In terms of energy, the first two syllables may be perceived as being relatively more even in the Split variety. It is quite likely that this relatively equal energy, combined with the high fall in the post-stress syllable for the rising prosodemes, is an important factor in the perception that the location of stress is not quite so obvious in the Split variety, particularly for the rising prosodemes. As such, speakers may perceive neither the fully retracted Neo-Štokavian *Màrija*, nor the unretracted Čakavian *Marĭja*, but an intermediate double accent form: *Màrĭja.*

At this stage, we may tentatively hypothesize that historical stress retraction does not involve the simultaneous retraction of all cues to stress. In the present case, it seems that whilst durational cues retracted, energy cues may not have – or alternatively, energy cues were not present, but were subsequently innovated when duration cues retracted. Importantly, the High tone that did not retract (i.e. the High tone of the rising prosodeme) seems to be extra high in the Split variety, leading to the perception of a fall in the originally stressed syllable. In the standard variety, the High tone of the rising prosodeme is not quite as high: this suggests that the retraction of stress later led to a more salient pitch movement (i.e. a rise) on the newly stressed syllable.

5. Acknowledgements

We would like to thank our speakers for their time and dedication to language research (Ena Marinković, Jelena Pocedulić, Josipa Papeš, Josipa Teskera, Magdalena Andromak, Nuša Vrdoljak, and two other speakers who wished to remain anonymous), and Jordan Bićanić for help with the recordings. The ethics protocols for this work were approved by the Office of Research at the University of California, Santa Barbara. Financial support was provided by the School of Humanities and Social Sciences and the Centre for Research on Language Diversity at La Trobe University.

6. References

[1] Holzer, G., Untersuchungen Zum Urslavischen Einleitende Kapitel, Lautlehre, Morphematik. Frankfurt a.M.: Peter Lang GmbH, Internationaler Verlag der Wissenschaften, 2020.

[2] Kapović, M., "The Unattainable Standard – Zagreb Dialect Meets Standard Croatian Accentuation", *Slověne. International Journal of Slavic Studies,* **1**(7):337-36, 2018.

[3] Magner, T. F. "City Dialects in Yugoslavia", American Contributions to the Eighth International Congress of Slavists (Zagreb and Ljubljana, September 3-9, 1978). Ohio: Slavica Publishers, Inc., 465–482, 1978.

[4] Kapović, M., Povijest hrvatske akcentuacije. Fonetika, Matica hrvatska, Zagreb, 2015.

[5] Godjevac, S., "Transcribing Serbo-Croatian intonation", in S-A. Jun [Ed], Prosodic Typology, 146-171, Oxford, 2005.

[6] Browne, E. W. and McCawley, J. D., "Srpskohrvatski akcenat", Zbornik za filologiju i lingvistiku VIII:147–151, 1965.

[7] Inkelas, S. and Zec, D. "Serbo-Croatian pitch accent: the interaction of tone, stress, and intonation", Language, 64:227-248, 1988.

[8] Rešetar, M., Die serbokroatische Betonung südwestlicher Mundarten, Alfred Hölder, K. u K. Hof- und Universitäts-Buchhandler, Wien, 1900.

[9] Ohala, J., "Production of tone", in V. Fromkin [Ed], Tone: A Linguistic Survey, 5-39, Academic Press, 1978.

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

[11] Winkelmann, R., Harrington, J. and Jänsch K. EMU-SDMS: Advanced speech database management and analysis in R. *Computer Speech and Language* 45 392-410, 2017.

[12] Winkelmann, Raphael, Jaensch, K., Cassidy, S. and Harrington, J. emuR: Main Package of the EMU Speech Database Management System. R package version 2.0.4, 2019.

[13] R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/, 2020.

[14] Sjölander, K. Snack Sound Toolkit, Stockholm: KTH Royal Institute of Technology, Retrieved from http://www.speech.kth.se/snack. 2014.

[15] Vicenik, C., Lin, S., Keating, P. and Shue, Y-L. Online documentation for VoiceSauce. Available at http://www.phonetics.ucla.edu/voicesauce/documentation/index.html, 2020

[16] Wickham, H. ggplot2: Elegant Graphics for Data Analysis, Springer-Verlag New York, 2016.

[17] Academy Grammar (AG) = S. Babić, D. Brozović, M. Moguš, S. Pavešić, I. Škarić, S Težak, Povijesni pregled, glasovi i oblici hrvatskoga književnog jezika. Zagreb : Hrvatska akademija znanosti i umjetnosti–Globus, 1991.

A corpus study of word (root) prominence in Vera'a

Catalina Torres, Stefan Schnell

University of Zurich
ARC Centre of Excellence for the Dynamics of Language
catalina.torres@ivs.uzh.ch, stefan.schnell@uzh.ch

Abstract

This study presents an acoustic investigation of word level prosody in the Oceanic language Vera'a. The analysis is based on a corpus of speech data collected during fieldwork from multiple speakers. A previous description of Vera'a suggests the language has lexical stress but its acoustic realisation was not further investigated. This study provides the first instrumental examination of five acoustic measures and their relation to prominence marking. The evidence indicates that vowels in the last syllable of the root are more prominent. However, these observations are restricted to some acoustic parameters.
Index Terms: root, word prominence, acoustic correlates, Oceanic languages, prosodic typology

1. Introduction

Despite a growing body of research dealing with the phonetics and phonology of word prosodic systems in the Austronesian language family [1, 2, 3, 4, 5, 6, 7, 8], instrumental quantitative studies in this field remain under-represented. Word prosodic systems in the Austronesian language family are of particular interest since previously reported lexical stress patterns have been challenged in the literature [9] showing a tendency for phrasal edge-marking prosodic typology [10]. As an example, in Indonesian, a language often spoken by multilingual speakers, studies have shown that lexical stress might not be present. Instead the evidence suggests a pattern of phrasal prominence. Additionally, these studies show that perception and production of prominence may vary depending on other languages spoken by individuals [11, 12].

As Gordon and Roettger [13] point out, one particular issue that studies examining word prosodic systems struggle with is a lack of control for prosodic boundary phenomena. The authors suggest to control for target words not to be placed at phrasal boundaries and to include a sufficiently large sample of lexical items (although they do not specify number of items or speakers). Efforts to implement state of the art phonetic analysis on under-documented languages from the Pacific have yielded important results. Recent instrumental work on Austronesian languages from the Oceanic subgroup indicate that impressionistic reports from the literature do not bear out when tested experimentally. These studies include data from multiple speakers and tested speech produced in different contexts while controlling for phrasal boundary effects. Contra previous claims, acoustic phonetic studies on Nafsan and Drehu show that these are edge-marking languages and prominence is realised post-lexically [5, 14, 8]. Interestingly, Nafsan and Drehu are languages that do not display weight sensitivity in their prominence marking systems. Although in both languages there is a phonemic vowel length distinction [15, 16, 17], prominence lending pitch peaks align with boundaries of phrases [5] or prosodic constituents such as accentual phrases [8]. We here report findings from a corpus study of Vera'a, another Oceanic language that is closely related to Nafsan and Drehu within the subgroup of Eastern Oceanic languages and is spoken in North Vanuatu.

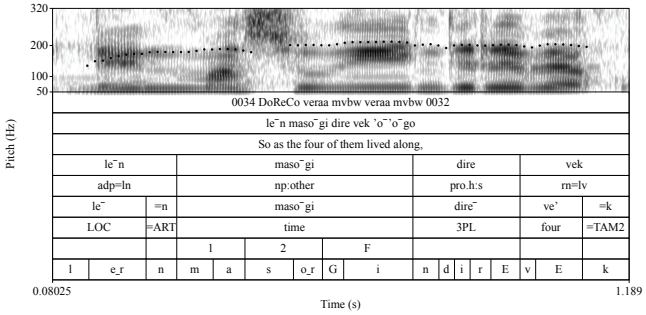


Figure 1: *Waveform, F0 trace, and spectrogram of an utterance from the corpus. The root in the utterance is masogi 'time'.*

1.1. Typological features of Vera'a

Vera'a is a canonical Oceanic language in the sense of [18]: Phrase structure is head-initial and word order on clause level is SVO. Adpositional phrases are all prepositional phrases, and on NP level heads precede any modifiers. The only exception is that articles precede NP heads, and so does a small set of quantifiers. The language is mostly analytic, the only affixes being possessive suffixes that attach to nominal roots. The phonology of Vera'a is similar to that of nearby languages: it has 17 consonant phonemes – including a double-articulated labio-velar plosive and a double-articulated labio-velar nasals – and seven vowel phonemes (/ɐ, e, ɛ, i, o, ɔ, u/). Vera'a differs from a range of languages in its neighbourhood though in that vowel length is not phonemic. Syllable structure is (C)V(C), occurring word internally and finally. Complex margins of the type CCV or VCC are only marginally present. In the sketch grammar of the language [19], word stress was impressionistically determined to fall on the final or penultimate syllable of a word (a word equals a bare root in our study). Alternatively, when the last syllable of a word is heavy (e.g., CVC) it is said to bear stress. There is no indication that stress could be distinctive in Vera'a and it is not known that words are distinguished based on stress.

1.2. Methodological considerations

In this exploratory study we focus on selected acoustic parameters that have been associated with stress and prominence marking cross-linguistically [20]. Taking into account recent evidence from other Oceanic languages [5, 14, 8] this study seeks

to remain agnostic as to whether there is lexical stress in the language, especially considering that there is no evidence of minimal pairs based on this criterion. Instead we set out to examine whether there is acoustic evidence for a prominence pattern at the word/root level. In doing so we leave the question open as to whether Vera'a displays a lexical stress-accent to which post-lexical phrase accents dock on, as typically known from languages such as English [13, 21]. As previous studies have shown, some languages can also display misalignment between stress and post-lexical phrase level prosody meaning that a demarcating and highlighting function need not be aligned [6, 22]. A more detailed explanation for this methodological choice can be found in the discussion. Similar to [4], we draw on narrative discourse data which has the advantage of representing speech production in connected discourse rather than isolated utterances. Our data, however, stems from a larger language documentation corpus, which bears the additional advantage of these data to be representative of the common speech production in everyday language use.

2. Research aims

Considering recent findings on word and phrase level prominence patterns in Austronesian and Oceanic languages, it is of interest to examine Vera'a acoustically and determine whether there is phonetic evidence to support the previous description [19]. In particular, it is of interest to evaluate whether the penultimate or final syllable of roots, all other things being equal, represent the most prominent syllable. Additionally, it is our aim to examine whether there is evidence for heavy root-final syllables (CVC) to be acoustically more salient than heavy syllables in other positions. We focus on nominal roots in this paper for practical reasons.

3. Materials and Method

3.1. Corpus

Our corpus consists of 7 narrative texts, each produced by a single speaker primarily for the purpose of being recorded and added to the ongoing documentation of the Vera'a language (Schnell et al ongoing) as part of the local 'orature'. Speakers came from both genders (2f, 5m) with age ranging between 16 and approximately 65 at time of recording in early 2007.[1] These narratives constitute the Vera'a subcorpus of DoReCo [23, 24] on whose time-aligned segmentation the current study is based, and it is also a major part of the Vera'a corpus within Multi-CAST [25, 26] which features relevant morphosyntactic annotations that enable us to identify nominal roots which are the focus of this study. A detailed description of forced alignment and segmentation process such as location of root and vowel boundaries can be found in [27]. All narratives are in turn part of the larger Vera'a documentation corpus archived with The Language Archive (MPI Nijmegen) (`https://hdl.handle.net/1839/bc035bf8-1d9b-4163-8131-983d5a7b08ab`).

Recordings used in the current study were made as part of the general language documentation project of Vera'a. All recordings were made in bamboo-walled houses that keeps interference from wind and other background noises (community activity, animals, etc.) as minimal as possible in the given conditions. Speakers offered to be recorded as contribution to the documentation project.

[1] Exact age is often unknown by elderly speakers of the language.

3.2. Data curation

For the purpose of this study a selection of nominal word roots was carried out, similar to [4]. As NPs in Vera'a can feature both pre- and post-nominal satellites and nominal roots can be expanded by suffixes, the decision was made to base the analysis on forms that represent NP heads and are simple roots. First, the narratives in the corpus were segmented into phrases the size of short utterances. This initial corpus amounts to 1484 utterances. NPs that were not produced directly at the start or the offset of the utterance and that did not co-occur with pauses were identified. For each root an additional segmentation of syllables was included, following [19]. Additionally, the number of syllables per root and the syllable structure of all syllables in the corpus was marked. A hierarchical database was constructed using the EMU Speech Database Management System [28]. It included the following eight tiers: phonemic segments, syllables, morphological glossing, morphologically segmented Vera'a text, syntactic glossing, corresponding Vera'a text, English translation, Vera'a original utterance. Duration and acoustic values were queried using the emuR package in R [29, 30].

3.3. Analysis

Following the criteria described above, a set of 412 noun roots and 134 unique word-roots was identified. Table 1 shows an overview of the vowels, syllable structures, and number of syllables per root present in the set. Importantly, the roots selected were never placed at the onset or offset of the utterance. This was done with the aim to control that they were not at a major prosodic boundary. See example in Figure 1. The data set is not balanced containing words of different lengths and with different syllable structures. To allow a comparison across roots and control for potential confounding factors the following analyses are restricted to a set of CV (463 tokens) and another of CVC (280 tokens) syllables. Five acoustic measurements were taken for the vowels of the selected syllables: vowel duration, intensity at mid point, relative intensity, F1, and F2 at mid point. Relative intensity was calculated as the difference between intensity at vowel mid point and intensity at mid point in the last segment preceding the root. Raw values for F1, F2, and vowel duration were Lobanov normalised and these values were used in the statistical analyses.

Table 1: *Summary and counts of vowels, syllable structures, and number of syllables per root found in our corpus.*

								Total
Vowel	**ɐ**	**e**	**ɛ**	**i**	**o**	**ɔ**	**u**	
Count	284	94	137	82	75	58	57	787
Syllable structure			**cv**	**cvc**	**ccvc**	**v**	**vc**	
Count			463	280	1	35	8	787
Syllables per root			**1**	**2**	**3**	**4**	**5**	
Count			126	201	55	24	3	412

3.3.1. *Statistical analyses*

Data were analysed using linear mixed effects models. Statistical analyses were carried out in R [30] with help of the statistics package lme4 [31]. Values were fitted into a linear mixed effects model to investigate specific factors of interest. Following [32], no random slopes are added to the models as this affects statistical power of small data sets, such as the one used in this study.

4. Results

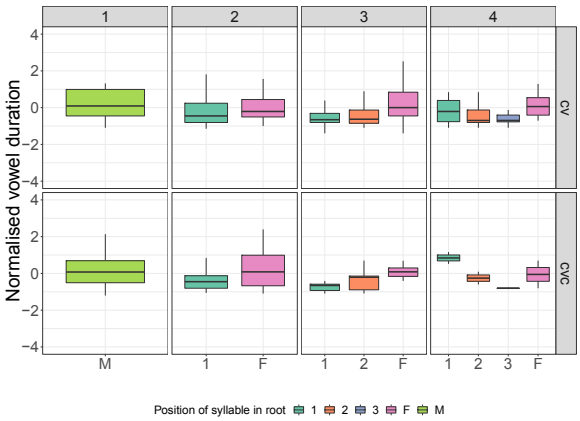


Figure 2: *Normalised vowel duration for CV- and CVC-tokens. Here, 1 stands for first syllable in the root, 2 for second, 3 for third, 4 for fourth, F for final and M for monosyllabic.*

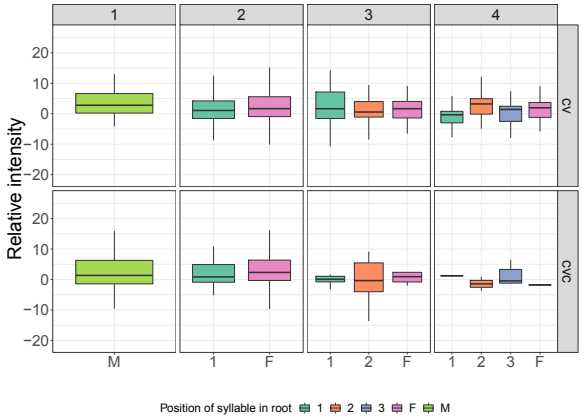


Figure 3: *Relative intensity in mono- and polysyllabic roots in CV and CVC-syllables.*

4.1. Duration

Figure 2 shows the normalised vowel duration for vowels obtained in words containing one to four syllables in CV and CVC structures. A model that had Position of syllable in the root, Syllables per root, Vowel quality, Sex, and Syllable structure as fixed factors together with speaker and word as random intercepts was performed to identify whether syllable position affects vowel duration. The results show that vowels in root final syllables (Est. 16 $\pm$ 2.6 ms, $p<0.0001$)[2] and in monosyllabic roots (Est. 20 $\pm$ 6.5 ms, $p<0.0001$) are significantly longer than vowels in other positions. The factor Syllable structure was not significant ($p=0.8$), showing that vowels in CV and CVC structure showed comparable duration. Separate models for the CV and CVC sets were additionally performed, but the overall result does not change, apart from estimates being larger in the CVC set, e.g for the final syllable (Est. 25 $\pm$ 5.3 ms, $p<0.0001$).

4.2. Intensity

Intensity was taken at vowel mid point to examine vowels as a function of their quality and position in the root. A model with Position of syllable in the root, Syllables per root, Vowel quality, Sex, and Syllable structure as fixed factors together with speaker and word as random intercepts was used to identify which factors affect intensity in the vowel. Similar to results in 4.1, the vowels in final position (Est. 1.3 $\pm$ 0.4 dB, $p<0.0001$) and monosyllabic roots (Est. 1.7 $\pm$ 0.9 dB, $p<0.05$) show significantly greater intensity. Additionally, an effect of syllable structure was found (Est. 0.9 $\pm$ 0.4 dB, $p<0.02$) with monosyllables in CVC showing lower intensity. The differences are, however, fairly modest.

4.3. Relative intensity

As noted in other studies examining intensity in relation to prominence [4, 7], it is helpful to evaluate relative intensity as this can provide a more reliable measure than raw values. For this purpose we calculated the difference between the intensity at vowel mid point and that of the last segment preceding the root analysed. Figure 3 shows the relative intensity according to position of the syllable in the root and syllable structure. The median values in the graph suggest that the final syllable displays greater relative intensity. We employed a model with Position of syllable in the root, Syllables per root, Vowel quality, Sex, and Syllable structure as fixed factors together with speaker and word as random intercepts. In line with results in 4.1 and 4.2, the vowels in final position display a significantly greater relative intensity (Est. 1.3 $\pm$ 0.4 relative intensity, $p<0.004$). In this case there was no significant result for monosyllables or syllable structure.

4.4. First and second formant

Figure 4 summarises the Lobanov normalised F1 for all Vera'a vowels in the first and second syllable of disyllabic words with CV. The median F1 of vowels /ɐ, e, o/ suggests these vowels have a raised F1 when in final position. However, this is not the case for the rest of the vowel inventory /ɔ, u ɛ, i/. Figure 5 shows the F1 trajectory of ɐ-vowels as produced by male speakers. The vowel /ɐ/ was selected because it is the vowel with the largest number of tokens (see Table 1). The trajectories indicate that the vowel appears to be realised as more peripheral in its F1 when in the final syllable or in a monosyllabic root. To investigate F1 and F2 at mid point two models with Position of syllable in the root, Syllables per root, Vowel quality, Sex, and Syllable structure as fixed factors together with speaker and word as random intercepts were used. The models included all seven vowel qualities present in our corpus. As can be expected the factor vowel quality was significant for all vowels. However, the factor of main interest to determine prominence patterns in the root, namely Position of syllable in the root did not yield any significant result for F1 or F2. Neither did we find a significant result for Syllable structure and separately run models for the CV and CVC sets do not show other results. Note however that the number of tokens per vowel varies greatly and that there might be too few vowels in all possible positions to determine whether the position in the root truly plays a role. Future investigations of Vera'a prominence patterns should include more vowel tokens.

[2]Estimates are provided in ms for a better overview.

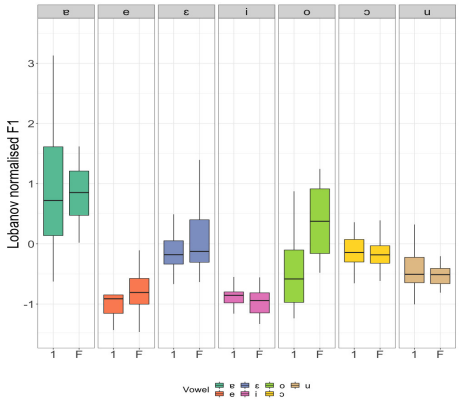


Figure 4: *Normalised F1 of disyllabic roots with CV-syllables.*

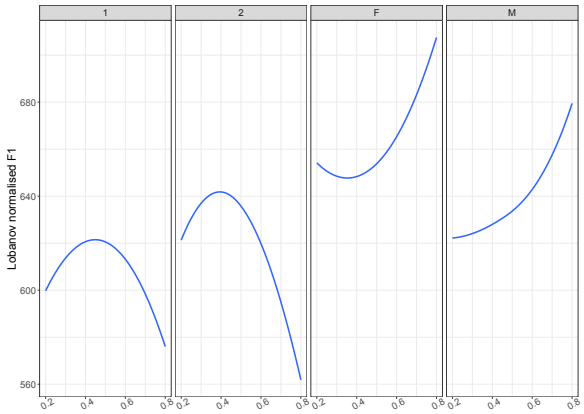


Figure 5: *Smoothed first formant trajectories based on 10 points between 0.20% and 0.80% of ɐ-vowels in CV. Only vowels that were preceded or followed by plosive and fricative consonants are included. Tokens were produced by male speakers.*

5. Discussion and Conclusion

Based on a speech corpus of Vera'a narratives this study examined five acoustic parameters in relation to prominence marking. To this end noun roots consisting of mono- and polysyllabic words were compared. Our analysis indicates that the final syllable (CV and CVC) of polysyllabic Vera'a roots is more salient than others through means of duration and relative intensity. More precisely we found that vowels in CV and CVC root final syllables and in monosyllables, display significantly longer duration values than other syllables in the root. Additionally, it was established that relative intensity is greater in root final syllables. Since in monosyllabic roots prominence can only be marked on this syllable it is not surprising that final syllables and monosyllables show similar results regarding duration. In addition, this study also presents results on the acoustic measures raw intensity, normalised F1, and normalised F2. It was found that intensity is greater in vowels in root final syllables. Moreover, vowels in CV-syllables display greater intensity than those in CVC. The differences found in intensity appear to be of small magnitude and although they provide additional evidence in favour of a more prominent root final syllable, they should be interpreted with caution.

Figures 4 and 5 suggest that F1 is raised in the vowels /ɐ, e, o/ when in syllables in root final position. However, the statistical analysis did not confirm this observation. There are different factors that could have influenced this result. Note that a survey on acoustic cues to stress [20] found that in languages in which stress affects vowel quality, the effect is often limited to certain vowels and/or one formant. Additionally, due to the nature of the speech present in our corpus, this study does not contain an even number of vowel tokens in every possible syllable structure (CV and CVC) or position of the syllables (first, second, third, fourth, final, monosyllabic), this means that some vowels are not present in all positions (e.g., the vowel /ɛ/ in the third syllable in CV, or vowel /i/ in the first and third position in CVC). This shortcoming is due to the lower frequency of some vowels in these positions in our corpus but could be mitigated in future research by including a larger set of roots. No statistical evidence was found for F2 as a correlate of prominence in our corpus.

This study presents the first acoustic evidence for word/root prominence being marked on the last syllable in Vera'a, confirming some observations from [19]. In other words, it was found that root final syllables and heavy root final syllables are acoustically more salient than syllables in other positions, although similarly salient compared to monosyllables. Note however that no evidence was found for penultimate stress, contra the previous impressionistic report.

One aspect that was not examined in this study is the effect of fundamental frequency (F0). This was omitted for several reasons. First, to date there is no description of Vera'a phrase level prosody making it difficult to make any well-informed predictions. Second, cross-linguistically, it is known that F0 is often associated with phrase level prosody, marking not only word level but also post-lexical prominence, as well prosodic boundaries [33, 13, 20]. Although this study sought to control for effects relative to prosodic boundary marking of major intonation breaks, it is possible that other factors related to prosodic phrasing could have an effect on the roots investigated. Moreover, there is too little evidence to date to make predictions on Vera'a being a bottom-up language like English, or a top-down language such as Chickasaw [13]. This means that we could not reliably predict whether in Vera'a pitch accents would dock on to a potentially stressed syllable, or whether stress would be orthogonal to phrasal accent assignment, displaying word-level unstressed syllables carrying tonal events like Chickasaw [22] or potentially display a yet unknown alignment of F0. For these reasons we have restricted our analysis to the five acoustic measures presented above.

In conclusion, we find evidence that root final syllables are acoustically salient in Vera'a simple roots. This finding could be indicative of regular demarcative right edge prominence. A more detailed account of Vera'a prosodic typology needs to be completed to be able to determine whether the language conforms to patterns observed in more closely related languages with post-lexical prominence marking [5, 14, 8], whether it could show similarities to languages that display misalignment of F0 in highlighting and demarcating [6, 22] or whether it displays stress-accent [21]. This study is part of a larger research project examining the acoustics of Vera'a and prominence patterns in the language. Future work including a larger set of data and extending to examine the acoustic correlates of phrase level prosodic structure is planned.

6. Acknowledgements

We thank the speakers who participated, as well as Sabrina Ryffel and Pelin Teberoğlu for supporting data processing.

7. References

[1] R. Maskikit-Essed and C. Gussenhoven, "No stress, no pitch accent, no prosodic focus: The case of Ambonese Malay," *Phonology*, vol. 33, no. 2, pp. 353–389, 2016.

[2] R. Billington, J. Fletcher, N. Thieberger, and B. Volchok, "Acoustic correlates of prominence in Nafsan," in *Proceedings of the 17th Australasian International Speech Science and Technology Conference*, 2018, pp. 137–140.

[3] C. Torres, J. Fletcher, and G. Wigglesworth, "Investigating word prominence in Drehu," in *Proceedings of the 17th Australasian International Speech Science and Technology Conference*, 2018, pp. 141–144.

[4] C. Kaland, "Acoustic correlates of word stress in papuan malay," *Journal of Phonetics*, vol. 74, pp. 55–74, 2019.

[5] J. Fletcher, R. Billington, and N. Thieberger, "Prosodic marking of focus in Nafsan," in *Proceedings of the 19th International Congress of Phonetic Sciences, Melbourne*, 2019, pp. 1758–1764.

[6] C. Kaland and S. Baumann, "Demarcating and highlighting in Papuan Malay phrase prosody," *The Journal of the Acoustical Society of America*, vol. 147, no. 4, pp. 2974–2988, 2020.

[7] R. Billington, J. Fletcher, N. Thieberger, and B. Volchok, "Acoustic evidence for right-edge prominence in Nafsan," *The Journal of the Acoustical Society of America*, vol. 147, no. 4, pp. 2829–2844, 2020.

[8] C. Torres and J. Fletcher, "Phrase-level and edge marking in drehu," *Glossa: a journal of general linguistics*, p. 1–32, 2022.

[9] N. Himmelmann and D. Kaufman, "Austronesia," in *The Oxford Handbook of Language Prosody*, 2020.

[10] S.-A. Jun, "Prosodic typology: By prominence type, word prosody, and macro-rhythm," in *Prosodic typology II: The phonology of intonation and phrasing*, jun, sun-ah ed. Oxford University Press, 2014, pp. 520–539.

[11] E. van Zanten and V. J. van Heuven, "Word stress in Indonesian: Its communicative relevance," *Bijdragen tot de Taal-, Land-en Volkenkunde*, vol. 154, pp. 129–149, 1998.

[12] R. Goedemans and E. van Zanten, "Stress and accent in Indonesian," in *LOT Occasional series*. LOT, Netherlands Graduate School of Linguistics, 2007, vol. 9, pp. 35–62.

[13] T. Roettger and M. Gordon, "Methodological issues in the study of word stress correlates," *Linguistics Vanguard*, vol. 3, no. 1, 2017.

[14] C. Torres and J. Fletcher, "The alignment of F0 tonal targets under changes in speech rate in Drehu," *The Journal of the Acoustical Society of America*, vol. 147, no. 4, pp. 2947–2958, 2020.

[15] D. T. Tryon, *Dehu grammar*. Australian National University, 1968.

[16] C. Moyse-Faurie, *Le drehu, langue de Lifou (Iles Loyauté). Phonologie, morphologie, syntaxe*. Langues et Cultures du Pacifique Ivry, 1983.

[17] R. Billington, N. Thieberger, and J. Fletcher, "Nafsan," *Journal of the International Phonetic Association*, pp. 1–21, 2021.

[18] M. D. Ross, "The morphosyntactic typology of Oceanic languages," *Language and Linguistics*, vol. 5, no. 2, pp. 491–541, 2004.

[19] S. Schnell, "A grammar of Vera'a," Ph.D. dissertation, Kiel University, 2011.

[20] M. Gordon and T. Roettger, "Acoustic correlates of word stress: A cross-linguistic survey," *Linguistics Vanguard*, vol. 3, no. 1, 2017.

[21] L. M. Hyman, "Do all languages have word accent," in *Word stress: Theoretical and typological issues*, H. van der Hulst, Ed. Cambridge University Press Cambridge, 2014, pp. 56–82.

[22] M. Gordon, "The phonology of pitch accents in Chickasaw," *Phonology*, vol. 20, no. 2, pp. 173–218, 2003.

[23] S. Schnell, "Vera'a doreco dataset," in *Language Documentation Reference Corpus (DoReCo) 1.1*, F. Seifart, L. Paschen, and M. Stave, Eds. Berlin & Lyon: Leibniz-Zentrum Allgemeine Sprachwissenschaft & laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), 2022. [Online]. Available: https://doreco.huma-num.fr/languages/vera1241

[24] F. Seifart, L. Paschen, and M. Stave, Eds., *Language Documentation Reference Corpus (DoReCo)*. Leibniz-Zentrum Allgemeine Sprachwissenschaft and Laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), 2022. [Online]. Available: https://doreco.huma-num.fr/

[25] S. Schnell, "Multi-CAST Vera'a," in *Multi-CAST*, G. Haig and S. Schnell, Eds., 2015, version 2207.

[26] G. Haig and S. Schnell, Eds., *Multi-CAST*, 2015. [Online]. Available: https://multicast.aspra.uni-bamberg.de/

[27] L. Paschen, S. Fuchs, and F. Seifart, "Final lengthening and vowel length in 25 languages," *Journal of Phonetics*, vol. 94, p. 101179, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0095447022000547

[28] R. Winkelmann, J. Harrington, and K. Jänsch, "Emusdms: Advanced speech database management and analysis in r," *Computer Speech & Language*, vol. 45, pp. 392–410, 2017.

[29] R. Winkelmann, K. Jaensch, S. Cassidy, and J. Harrington, *emuR: Main Package of the EMU Speech Database Management System*, 2017, r package version 0.2.3.

[30] 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/

[31] D. Bates, M. Mächler, and S. Walker, *Linear Mixed-Effects Models using 'Eigen' and S4*, 2019.

[32] H. Matuschek, R. Kliegl, S. Vasishth, H. Baayen, and D. Bates, "Balancing Type I error and power in linear mixed models," *Journal of memory and language*, vol. 94, pp. 305–315, 2017.

[33] D. R. Ladd, *Intonational phonology*. Cambridge University Press, 2008.

Duration in Zhangzhou Southern Min: Variation, Correlation, and Constraint

Yishan Huang

University of Sydney; Australian National University

yishan.huang@sydney.edu.au

Abstract

This study explores the status of duration in Zhangzhou Southern Min with rich tonal contrasts across three different linguistic contexts. Several innovative findings have gone beyond conventional assumptions and advanced our knowledge of how suprasegmentals (F0 and duration), segments (coda types), and linguistic contexts (citation, phrase-initial, phrase-final) interact to shape dynamic realisations of duration in the tonal phonology and morphosyntax of Zhangzhou speech. It is hoped to contribute valuable empirical data to the typology of tonal studies in Asia and shed important light on how humans encode and decode duration in their cognitive grammar.

Index Terms: duration, F0, tone, contexts, Zhangzhou, Southern Min

1. Introduction

As a linguistic phenomenon, Duration has been conventionally viewed as an important property of speech segments to convey linguistic content and/or paralinguistic information, but languages use duration differently. Some employ duration for segment intelligibility. For example, in English, the five tense vowels /i, e, ɑ, o, u/ are found longer than the six tense vowels /ɪ, ɛ, æ, ɔ, ʊ, ʌ/ by 38 ms [1]; segment types can be ranked hierarchically in terms of duration: vowels (135ms) > fricative (88ms) > nasals (87ms) > glides (70ms) > liquids (69ms) > stops (30ms), as cited in [1]. Contrastively, some other languages, though the proportion is barely about 3% of the world's languages, use duration for phonemicity, whereby different meanings of lexical items or grammatical structures can be distinguished [2]. For example, in Niger-Congo languages, a phonemic contrast exists between short and long vowels, along with many geminates indicating longer consonants in their segmental inventory [1].

Beyond the scope of segments, a great deal of attention has also been given to exploring the duration in the domain of the suprasegmental system, especially in those Asian languages with rich tonal contrasts. For example, the intrinsic correlation between duration and pitch has been studied cross-linguistically [3], [4], [5], and [6]. In general, vowels with low-pitched tones are considered to have a longer duration than with high-pitched tones, and vowels on rising tones appear longer than on falling or level tones [5] and [6]. In other words, pitch/F0 height and duration are correlated negatively: The higher the pitch/F0, the shorter the duration; the lower the pitch/F0, the longer the duration. However, counterexamples are also found to show a positive association between pitch/F0 and duration, such as in Taiwanese [3] and [4], Cantonese [6], and Yucatec Maya [6].

This study, built upon field linguistics, phonetics, phonology, and statistics, explores the duration status in the rich tonal contrasts of the southern min dialect of Zhangzhou in three different contexts: citation tone, phrase initial tone, and phrase-final tone. It is shown that tones can be classified into different groups in terms of duration, and the realisations can be alternated in accordance with the change of linguistic contexts, and so are the intrinsic relations between F0 and duration. It is guided by the following specific research questions: (a) What function does duration serve in the tonal system of Zhangzhou? (b) How are tones realised in duration across different linguistic contexts? and how many lengths are statistically different across contexts? (c) How are tonal duration and F0 related to each other? and how does their correlation change in accordance with the changing linguistic contexts? (d) Are tonal duration realisations affected by surrounding environments? If so, to what extent are they affected, and what conditions the variations? (e) how are durational realisations in the sandhi position related to their corresponding forms in the citation?

It is hoped to reveal how the local residents employ the parameter of duration as part of the suprasegmental system for communication purposes. It aims to advance our understanding of how suprasegmentals (F0 and duration), segments (coda types), and linguistic contexts (citation and disyllabic phrase) interact to shape dynamic realisations of duration in the tonal phonology and morphosyntax of Zhangzhou Southern Min. It is also hoped to contribute empirical data to the typology of tonal studies in Asia and shed important light on how humans encode duration in their cognitive grammar.

2. Research material and design

Zhangzhou is a prefecture-level city in the south of Fujian province of Southern China with a registered population of about 5.10 million. The colloquial language spoken by native Zhangzhou people is predominantly Southern Min, which is mutually intelligible with other Southern Min varieties (e.g., Taiwanese, Xiamen, and Quanzhou) but is entirely unintelligible with other Sinitic dialects (e.g., Mandarin, Hakka, Cantonese, Wu, Xiang, and Gan). The data used in this study were collected by the author in the urban districts of Longwen and Xiangcheng from 21 native speakers (9 males and 12 females) who were selected based on a set of strict criteria with an average age of 56.5 for males, and 50 for females [7] and[8].

Two corpora were used in this study. One is about 160 monosyllabic tokens for investigating citation tones with an average of 20 tokens for each tone. The other one is about 588 tokens for investigating tone sandhi behaviour across 64 (=8 tones * 8 tones) tonal combinations in disyllabic constructions, with an average of 12 tokens for each combination, but some combinations, in particular with tone 8, had less than 12 tokens to be processed, because the tone was less productive in this dialect. Tokens were elicited by individual speakers in Praat via a professional cardioid condenser microphone at a sampling frequency of 44100 Hz. Tonally relevant duration for each token incorporated those elements that excluded the syllable onset. Acoustically, the durational onset was set at the glottal

pulse where the amplitude of air pressure fluctuation began to increase; the periodicity of speech wave vibration appeared regular in the waveform, and the formant patterns in the spectrogram were clearly stable and identifiable. The offset was set at the point where periodicity and formant patterns cease to be visible. F0 and duration values were extracted using a script at ten equidistant sampling points in Praat.

Because acoustic signals are highly variable, the process of normalisation was applied to abstract away the variable content from the invariable linguistic content in this study, with the formula (1) z-score normalisation approach for F0 and (2) the absolute approach for the duration [7], [8] and [9]. For example, each tonal duration was expressed as a percentage of the average duration of all tones from the speaker being considered.

$$Z_i=(X_i-m)/s \qquad (1)$$

$$D_{norm}=(D/D_{mean})*100 \qquad (2)$$

Because this study involved 8 citation tones and 64 tonal combinations, the technique of pairwise t-test comparison by effect size [10] was employed to determine whether the variables (e.g., duration) among a set of tones (e.g., citation tones) differ from each other in a statistically significant way. For example, there were 28 (=8*7/2) paired differences to be tested in the citation context, as illustrated in Figure 1.

```
      tone1   tone2   tone3   tone4   tone5   tone6   tone7
tone2 8.6e-10 -       -       -       -       -       -
tone3 < 2e-16 < 2e-16 -       -       -       -       -
tone4 < 2e-16 < 2e-16 8.0e-06 -       -       -       -
tone5 2.3e-06 1.000   < 2e-16 < 2e-16 -       -       -
tone6 < 2e-16 < 2e-16 < 2e-16 < 2e-16 < 2e-16 -       -
tone7 < 2e-16 < 2e-16 < 2e-16 < 2e-16 < 2e-16 < 2e-16 -
tone8 1.1e-10 1.000   < 2e-16 < 2e-16 0.041   < 2e-16 < 2e-16
```

Figure 1: *Example of pairwise t-tests in citation context.*

The Bonferroni correction was applied to control for the Type I Error and achieve a significance. The corrected alpha was calculated by dividing the critical P value by the number of comparisons. For example, the corrected alpha was 0.00186 (= 0.05/28) in the citation. If the calculated t value was less than the corrected alpha, the paired difference was considered statistically significant, and vice versa. The testing result was visualised using the hierarchical clustering algorithm to help assess how many groups the sets of data can be clustered into from a scientific perspective.

3. Duration in Citation Form

Zhangzhou presents eight tones, although to fully appreciate this finding, one needs to recognise multidimensional characteristics of tonal realisations across different linguistic contexts [7] and [8]. Figure 2 shows the normalised F0 system of Zhangzhou citation tones, in which all F0 contours are expressed as a function of their corresponding normalised duration values from 21 speakers, representing the central tendency of this dialect as an independent variety. As indicated, tones in this dialect exhibit variation in both F0 contour shape and F0 height, which, on the one hand, involves four contour shapes of rising, level, mid-low level with a final fall, and falling contours, while, on the other hand, involving four contour heights of high, mid-high, mid, and low.

Like the F0, tones vary considerably in duration in this citation context. Figure 3 shows the acoustically normalised duration system of Zhangzhou in which each bar is expressed as a percentage of the average duration value of all tones from 21 speakers. Exhaustive pair-wise t-tests were conducted to compare 28 (8*7/2) paired differences under the assumption that all putative duration levels are independently and identically distributed. The testing results were visualised hierarchically, as shown at the bottom of figure 3. The threshold selected at 1 clusters the eight citation tones into four classes, with 1 representing the longest and 4 the shortest. Thus, a duration system of four levels can be achieved, which are noted as extra-long [V::], long [V:], medium [V], and short [V̆], as summarised in Table 1.

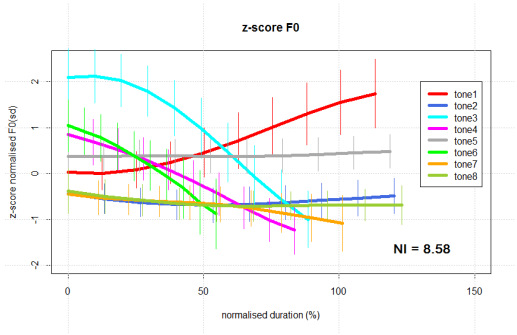


Figure 2: *Normalised F0 system of Zhangzhou citation tones.*

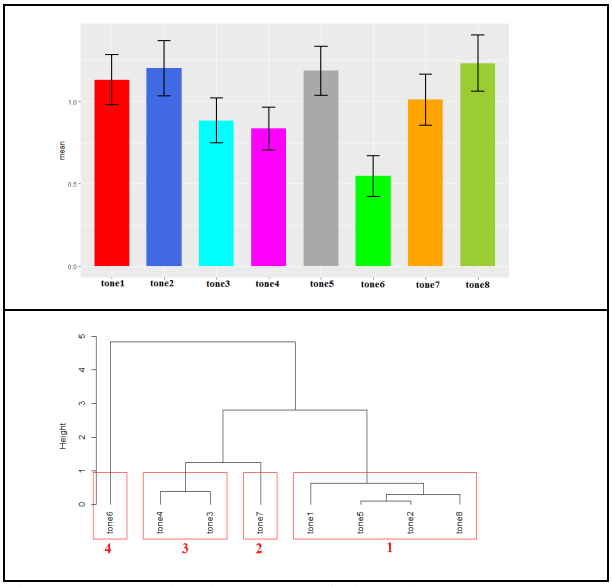


Figure 3: *Normalised Duration of Zhangzhou citation tones.*

Table 1. *F0 and duration of Zhangzhou citation tones.*

Tone	F0	Duration	Notation	Ranking
1 2 5 8	[35] [22] [33] [22]	extra long	[V::]	1
7 stopped	[221]	long	[V:]	2
3 4	[51] [41]	medium	[V]	3
6 stopped	[41]	short	[V̆]	4

Table 1 summaries the F0 and duration of these citation tones. Several aspects of linguistic significance can be observed with respect to the interaction between F0 and duration. (a) No statistically significant differences exist among the level (tones

2, 5, and 8), and among the unstopped falling tones (tones 3 and 4). In other words, the realisation of duration is not categorically affected by the changing F0 height, failing to support the across-linguistic assumption of a negative correlation between pitch height and duration. (b) The rising contour (tone 1) is grouped with the level contours (tones 2, 5, and 8), signifying no statistical hierarchy existing between rising and level-pitched tones in terms of duration. However, it is longer than any other falling contours, supporting the universal assumption that an upward F0 has a marked tendency to take a longer time than a downward change [11] and [12]. (c) The stopped tones are statistically significantly shorter than their corresponding unstopped tones that share a similar F0 realisation. For example, the stopped tone 6 shares a similar mid-high falling contour with the unstopped tone 4, but its duration is shorter. This indicates that the difference in syllable type can constrain the durational realisation because the stopped tones are associated with obstruent-ending syllables, while the unstopped tones with sonorant-ending syllables at the underlying level. (d) The stopped tones are not always the shortest as conventionally assumed in Sinitic studies [13], [14] and [15]. Instead, they can be longer than many other tones. For example, in citation, the stopped tone 7 is statistically significantly longer than tones of falling contours (tones 3, 4, and 6). (e) Tones 2 and 8 neutralise their contrasts in the citation context because of sharing common F0 and duration values. The reason why they are proposed as two different tones is because of their various realisations in other linguistic contexts, such as in the phrase-initial setting.

4. Duration in the Phrase-initial context

Zhangzhou presents a right-dominant sandhi system [7] and [8]. Tones at the non-rightmost position undergo categorical alternation but preserve their categories as their citation forms in the rightmost context. Its sandhi system is found not to be affected by the categories of surrounding tones; whereby individual tones are realised categorically the same regardless of which phonological environment they occupy; however, they may show a certain degree of phonetic variation due to co-articulation and position effects [7], and [8]. Figure 4 plots the acoustically normalised F0 system of phrase-initial tones comprising mid-level, rising and falling. Tones 2 and 3 have two statistically different variants in their offset values, which can be ascribed to the effect of regressive assimilation on the onset of their following tones. Tones 2 and 8 that share an identical F0 contour in citation can be realised differently in this context: tone 2 presents a mid-level, while tone 8 shows a mid-falling contour. As well as this, the whole F0 range is raised phrase-initially, with the average normalised values situating between -0.83 and 2.33, compared with the values between -1.44 and 2.1 in the citation context.

Figure 5 plots the normalised duration of the eight tones at the phrase-initial context across 64 (=8*8) tonal combinations, among which only tone 5 has two variants, with the value before tones 3 and 6 being statistically shorter. The nine phonetic duration levels form 36 (=9*8/2) paired differences to be tested by effective size. The result is visualised and shown at the bottom of figure 5. Thus, a system of two lengths can be achieved for the tones in this context, which are noted as medium [V] and extra-short [V̆] in Table 2. Additionally, several aspects deserve further discussion.

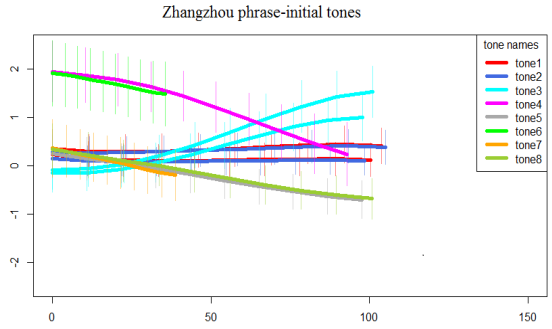


Figure 4: *F0 system of Zhangzhou phrase-initial tones.*

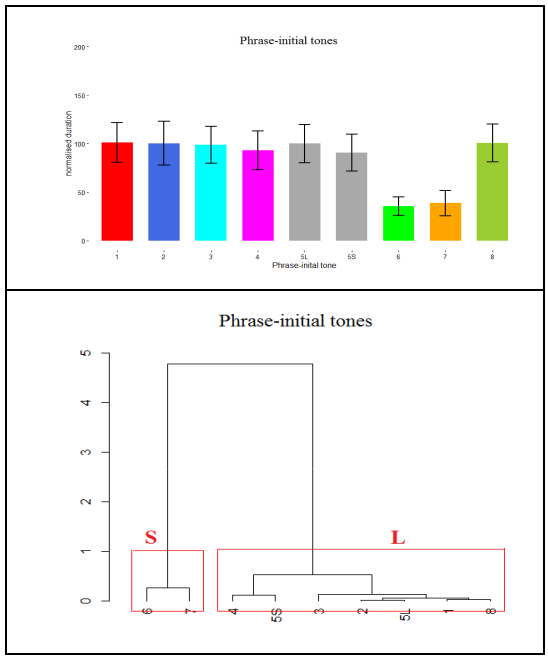


Figure 5: *Durations of Zhangzhou phrase-initial tones.*

Table 2. *F0 and duration of Zhangzhou phrase-initial tones*

Tone	F0	Duration	Notation
1	[33]	medium	[V]
2	[33]		
3	[24]/[35]		
4	[63]		
5	[32]		
8	[32]		
6 stopped	[65]	extra short	[V̆]
7 stopped	[32]		

(a) The number of lengths is considerably reduced phrase-initially. Only two lengths are tested statistically significantly different compared to four in the citation. (b) Phrase-initial tones are subject to neutralisation processes severely. The durational contrasts of the unstopped tones are neutralised to be medium. In contrast, that of stopped tones are neutralised to be extra short. (c) No significant difference exists between duration and F0. In other words, the durational realisation in this context is not affected by the F0 contour shapes and heights of surrounding tones at both phonological and phonetic levels. This can be seen from those six unstopped tones with variable

F0 contours of mid-level, low rising, extra-high falling, and mid falling have an identical medium duration. This manifestation not only falsifies the cross-linguistic assumption of a negative relation between F0 height and duration but also denies the universal tendency that considers a rising contour having a longer duration than falling one. (d) The conditioning factor on the two-length distinction appears to be solely associated with syllable structure. The medium length occurs on sonorant-ending syllables, regardless of tonal phonology and phonetics, whereas the extra short length occurs on the syllables ending in obstruent codas. (e) In this context, the parameter of duration can be seen as having a similar function as the syllable coda type that classifies syllables/tones into either stopped or unstopped categories. (f) A new duration value of an extra-short property emerges phrase initially, which can be regarded as a coupling effect of position (phrase-initial) and syllable type (obstruent-ending).

5. Duration in the Phrase-final context

The tones in the phrase-final context are presumed to maintain their categories and forms as their citation counterparts under the conventional assumption of a right-dominant sandhi system [16], [17], [18], [19], [20] and [21]. However, in Zhangzhou dialect, the right-most tonal realisations are highly sensitive to the phonetic environment of their preceding tones and presenting variation. This can be seen in figure 6, which plots the normalised F0 system of phrase-final tones across 64 tonal combinations. All tones except tone 3 have two variants, with the onset value statistically significantly higher after tones with a non-low F0 offset. Similarly, as seen in figure 7 about the normalised duration system, tones 5 and 7 are found to have two variants with one marginally longer than the other, resulting in 10 phonetic variants to be pairwise t-tested by an effect size of 45 (=10*9/2) paired differences, the result of which is visualised using the clustering algorithm. As seen, the phrase-final tones are clustered into three classes which can be labelled as extra-long [V::], long [V:], and medium [V], as summarised in Table 3. Tone 7 has a medium duration across most phrase-initial tones but has a long variant that only occurs after tone 8, which can thus be seen as a marked form.

Additionally, there also are several interesting aspects being further noted. (a) No statistically significant difference exists between rising and level contours, but both are longer than falling contours. (b) No statistically significant difference exists between low falling and high falling tones, but both are longer than the non-stopped tone 4 of a mid-high falling contour. (c) Tone 7 presents a shorter duration than tone 2, which shares similar F0 contours. (d) The number of duration differences is reduced to three (extra-long, long, and medium) from four (extra-long, long, medium, and short) in the citation. (e) The duration of most phrase-final tones is not the same as their corresponding citation form. Tones 2 and 8 have shorter durations than their citation forms, while tone 6 becomes longer in the phrase-final position. However, tones 1 and 5 maintain their extra-long property, while tone 4 keeps a medium level across citation and phrase-final contexts. This manifestation may be ascribed to the factor from the position effect and phonetic adjustment to preceding tones. (f) The durational manifestation of the phrase-final tones can also question the conventional assumption regarding the property of right-dominant tone sandhi because the forms of right-most tones do not necessarily preserve their corresponding citation forms without change.

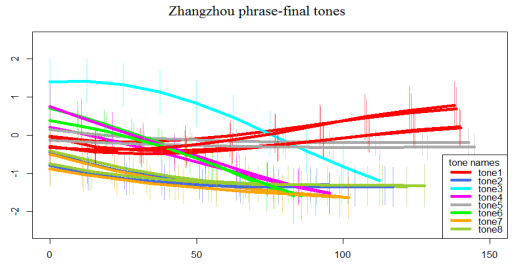


Figure 6: *F0 system of Zhangzhou phrase-final tones.*

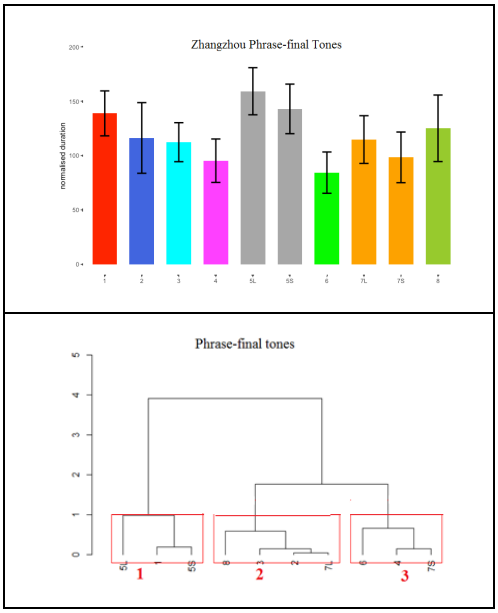


Figure 7: *Durations of Zhangzhou phrase-final tones*.

Table 3. *F0 and duration of Zhangzhou phrase-final tones.*

Tone	F0	Duration	Notation
1 5	[34]/[35] [33]/[43]	extra long	[V::]
2 3 8	[211]/[311] [52] [211]/[311]	long	[V:]
4 6 stopped 7 stopped	[41]/[51] [41]/[51] [211]/[311]	medium (marked long)	[V]

6. Conclusion

As discussed, the duration can function as F0 to distinguish Zhangzhou tones into different categories. The durational realisations are seen changing in accordance with the changing linguistic contexts and showing variation, which are also affected by the phonetics of surrounding tones and syllable types. This study fills in our knowledge of the Zhangzhou dialect but also upgrades our understanding of the relations between F0 and duration and the phonetic characteristics of a right-dominant tone sandhi system while falsifying several conventional assumptions. It contributes valuable empirical data to the typology of tonal studies in Asia and sheds important light on how humans encode and decode duration in their cognitive grammar.

7. Acknowledgements

I would like to express my sincere gratitude to Phil Rose, Mark Donohue, Paul Sidwell, and Shinichi Ishihara for their supervision on my tonal study at the ANU, to Siva Kalyan for his assistance in R, and to Michael Proctor for his valuable discussions on phonetics. I also appreciate my three anonymous reviewers for their time and very helpful comments, which I have tried to incorporate.

8. Reference

[1] Koffi, E., "A comprehensive review of the acoustic correlate of duration and its linguistic implications", Linguistic Portfolios, 10, 2021.

[2] Kubozono, H., "The phonetics and phonology of geminate consonants: Book review", Journal of the Acoustical Society of America, 142 (4), 1756, 2017.

[3] Zee, E., "Duration and intensity as correlates of F0", Journal of Phonetics, 6(3): 221-225, 1978.

[4] Kong, Q., " Influence of tones upon vowel duration in Cantonese", Language and Speech, 30: 387-399, 1987.

[5] Yu, A. C., "Tonal effects on perceived vowel duration", Laboratory Phonology,151-168, 2010.

[6] Faytak, M., & Yu, A. C., "A typological study of the interaction between level tones and duration", In E. Zee (Ed.), Proceedings of the International Congress of the Phonetic Sciences, XVII. International Congress of the Phonetic Sciences, 2011.

[7] Huang, Y., Tones in Zhangzhou: Pitch and beyond (PhD dissertation: Australian National University), 2018. Retrieved from https://openresearch-repository.anu.edu.au/handle/1885/144938, accessed on 19 June 2020.

[8] Huang, Y., Tones in Zhangzhou: Pitch and beyond, Cambridge, UK: Cambridge Scholar Publishing, 2020.

[9] Huang, Y., Donohue, M., Sidwell, P., & Rose, P., "Normalization of Zhangzhou citation tones", In C. Carignan, & M. Tyler (Eds.), Proceedings 16th Australasian International Conference on Speech Science & Technology, 217-220, Sydney, Australia: The Australian Speech Science & Technology Association,

[10] Levshina, N., How to do Linguistics with R: Data exploration and statistical analysis, Amsterdam, Netherlands: John Benjamins Publishing Company, 2015.

[11] Gandour, J., "On the interaction between tone and vowel length: Evidence from Thai dialects", Phonetica, 34: 54-65, 1977.

[12] Ohala, J. J., & Ewan, W. G., "Speed of pitch change", Journal of the Acoustical Society of America, 53, 345, 1972.

[13] Ma, C., Studies of Zhangzhou dialect 漳州方言研究, Hongkong: Zongheng Chubanshe 纵横出版社, 1994.

[14] Yang, X., Studies of tones and regional cultures of Zhangzhou dialect 漳州方言声调与地域文化研究, Beijing: Zhongguo Shehui Kexue Chubanshe 中国社会科学出版社, 2008.

]15] Guo, J., Zhangzhou Southern Min 漳州闽南方言. Zhangzhou: Zhangzhou Library 漳州图书馆, 2014.

[16] Wright, M. S., A metrical approach to tone sandhi in Chinese dialects (Doctoral dissertation, University of Massachusetts Amherst), 1983. Retrieved from http://scholarworks.umass.edu/dissertations/AAI8310348

[17] Shih, C.-L., The prosodic domain of tone sandhi in Chinese (Doctoral dissertation, University of California at San Diego), 1986. Retrieved from https://www.researchgate.net/publication/36071823_The_Prosodic_Domain_of_Tone_Sandhi_in_Chinese

[18] Ballard, W. L., "The history and development of tonal systems and tone alternations in South China (Vol. 22)", Study of Languages and Cultures of Asia and Africa: Monograph Series 22, 1988.

[19] Chen, M., Tone sandhi, Cambridge, England: Cambridge University Press, 2000.

[20] Zhang, J., "A directional asymmetry in Chinese tone sandhi systems", Journal of East Asian Linguistics, 16: 259-302, 2007.

[21] Rose, P., "Complexities of tonal realisation in a right-dominant Chinese Wu dialect—Disyllabic tone sandhi in a speaker from Wencheng", Journal of the South East Asian Linguistics Society, 9: 48-80, 2016.

Modeling Interaction between Tone and Phonation Type in the Northern Wu Dialect of Jinshan

Phil Rose#1, Tianle Yang*2*

*Independent researcher, Australia
#ANU Emeritus Faculty, Australia
[1] https://philjohnrose.net, [2] u6512077@alumni.anu.edu.au

Abstract

Impressionistic and acoustic data are presented for the seven tones of the Wu Chinese dialect of Jinshan 金山, where tone is much more than just pitch. The independence of extrinsic phonation type from syllable Onsets is exemplified, and it is argued using quantified tonatory parameters that phonation type determines tonal pitch, not *vice versa*. Command-response modeling is then used to factor tone into depression and tonal target components, which enable a more precise understanding of Jinshan tonological structure.

Index Terms: tonal acoustics, Wu dialects, phonation type, depression, command-response model.

1. Introduction

"Tone is seldom, if ever, a matter of pitch alone." Thus Eugenie Henderson, one of the pioneer descriptive phoneticians and linguists of South East Asian languages [1]. Of course, pitch is criterial for tone: the definition of a tone language is, after all, one in which pitch is part of the phonological representation of words [2 p.4, 3 p.229]. However there are many tone languages, especially in S.E. Asia, where tonal pitch is closely intertwined with other segmental and suprasegmental aspects of the syllable and word, and this paper describes the tones of one of them: the northern Wu dialect of Jinshan in southern Jiangsu province.

Jinshan belongs to the Tàihú-Sūhùjiā 太湖苏户嘉 sub-subgroup of Wu. Although close to Shanghai, the two varieties differ a lot, with Jinshan having more complex tones and tone sandhi. But it is a typical Wu dialect in its complex interaction between tonal pitch, phonation type, duration, vowel quality, syllable-structure and syllable Onsets. This paper aims to show how, with speech science and, some might say, speech technology, two of these components – tone and phonation type – can be quantitatively disentangled.

Jinshan has, we think, not been previously described. A comparison of Wu dialect descriptions of 33 sites in 1928 and again in 1992 [4, 5] shows that they changed considerably in this sixty year period, and more recent socio-phonetic findings on closely related Wu varieties [6] suggest that tonal change is accelerating in metropolitan areas due to urbanisation. In a sense, therefore, this description may also constitute a salvage operation.

2. Procedure

2.1. Informants, elicitation

Because of recent changes in the speech of younger Shanghai speakers [7], it was considered advisable to collect data from older Jinshan speakers, and so nine speakers over 60 years old were selected and recorded by the second author, who is a native Jinshan speaker (albeit a youngish one). Three of them are described here: two males and a female.

Informants were given the list of 453 basic words for exemplifying Chinese dialect lexicon in [8 pp.18-26] and asked to read out the equivalent Jinshan word. Some of the recordings may be listened to at [9].

The recordings were first phonetically transcribed, and then manually labeled in *Praat*. Transcription is an essential part of the process: it enables one to become familiar with a voice and note features of possible phonetic and/or phonological importance (to take an actual example from the recordings, between-speaker variation in the use of implosives as opposed to voiceless unaspirated stops).

Tone acoustics were quantified with the same method used in previous studies of Wu varieties, e.g. [10 11]. A wideband spectrogram was generated in *Praat*, together with its wave-form and superimposed F0. The token's tonally relevant F0 was then identified, extracted with a *Praat* script, and modeled in *R* by an 8th order polynomial. This enabled F0 values to be sampled from the polynomial F0 curve with a sufficiently high sampling frequency (at 10% points of the curve as well as 5% and 95%) to capture the details of its time-course. Phonation type was quantified with *VoiceSauce* [12]. Interaction between tonal pitch and phonation type was modeled with an extended version of Fujisaki's *command-response* model [13 - 15].

3. Results

3.1. Auditory analysis

Table 1. Examples of Jinshan tones on monosyllabic words

Tone name	Example
high fall	pɔ 包*wrap,* pɪã 冰*ice*, sɔ 烧 *burn*, lɔ 捞 *carry*, tsʰu 搓 *roll*
low rise-fall	bi̯ã̯ 平 *flat*, ʑɛ̯ 裁 *cut*, li̯ɛ̯ 晾 *to dry*, lɑ̯ 狼 *wolf*
high level	tsʰɛ *to rub*, sɔ 扫 *to sweep* tsɔ 早*early*
(delayed) mid rise	tsɔ 罩 *cover*, tsʰɛ 踩 *to trample,* tɔ 到 *arrive*, tʰɔ 套 *sheath,* su 漱 *to rinse*
(delayed) low rise	ɦʊɛ̯ *to throw*, ʑɛ̯ 站 *to stand*, mã̯ 问 *to ask*
short stopped high	səʔ 塞 *block*, pɑʔ 剥 *peel*, tʰəʔ 脱 *take off,* vəʔ 勿 *not*
short stopped low rise	ŋi̯ɛ̯ʔ 热 *hot*, ʑə̯ʔ 直 *straight,* və̯ʔ 活 *live*

Conventional auditory phonetic and phonological analysis showed our speakers have seven tones, which can be named

after their pitch features as follows: *high fall*, *low rise-fall*, *high level, mid rise*, *low rise, short stopped high* and *short stopped low rise*. Table 1 gives some examples.

3.2. Acoustic description and tonological structure

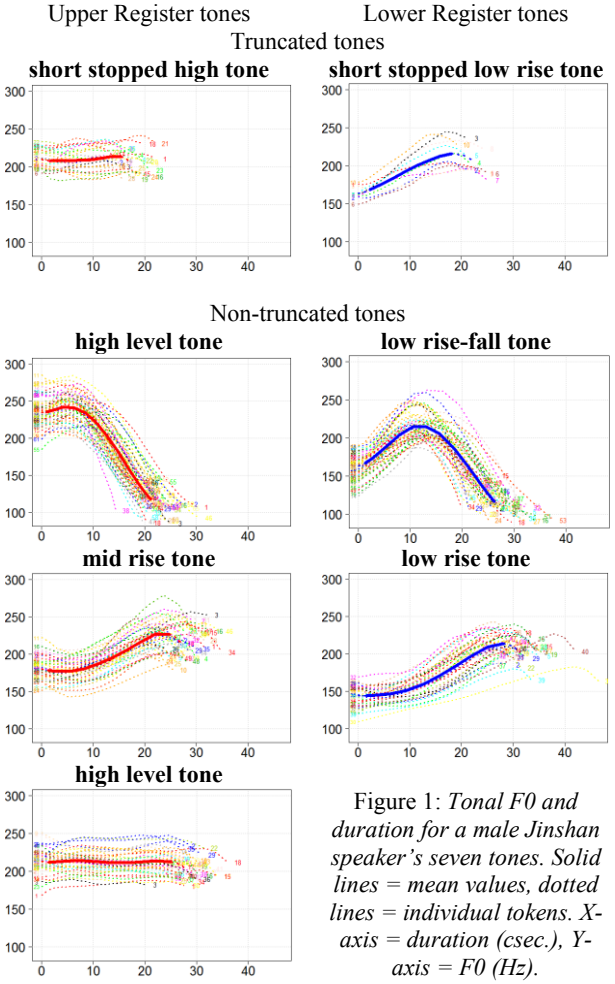


Figure 1: *Tonal F0 and duration for a male Jinshan speaker's seven tones. Solid lines = mean values, dotted lines = individual tokens. X-axis = duration (csec.), Y-axis = F0 (Hz).*

Figure 1 shows mean values for tonal F0, plotted as function of duration, for one of the male speakers. Individual tokens are also shown to give an idea of the amount of variation due to intrinsic factors. It can be seen that the tones' F0 shapes resemble their pitch descriptors fairly closely. It can also be seen that the short tones have about half to two-thirds of the duration of the long tones. The arrangement of the panels in figure 1 is important, as it shows how the seven tones are cross-classified in the typical Northern Wu manner by features of *truncation* and *register* [16]. Thus the truncated *short stopped high* and *short stopped low rise* tones have glottal-stop codas and shorter Rhymes, compared to the longer Rhymes, with gradual phonation offset, of the five other non-truncated tones.

For this paper, though, Register is the important dimension. Register partitions the tones into two sets. High fall, high level, mid rise and short high level belong to the upper register; low rise-fall, low rise and short low rise are lower register. Upper register tones are plotted on the left in figure 1, with their means in red; lower register tones on the right, with means in blue. Register correlates with many phonetic and phonological features. The most important of these, for this paper, are phonation type and depression (*depression* refers to the lowering of pitch onset as a function of phonetic and/or phonological factors producing, e.g., a rising-falling from a falling pitch [17 18]). The upper register tones have modal voice; the lower register tones have breathy voice and depressed pitch onset.

Register also correlates with the nature of Onset obstruents. Jinshan has the typical Wu three-way contrast within stop and affricate phonemes, e.g. /p^h t^h ts^h k^h, p t ts k, b d dz g/; and two-way contrast within fricatives, e.g. /f v, s z/. As in most Wu dialects, phonemic obstruent voicing is in complementary distribution with register: the voiceless phonemes co-occur with the upper register tones, and have the expected allophones e.g. /p/ → [p], /p^h/ → [p^h]. The voiced obstruent phonemes occur with the lower register tones and have different realisations conditioned by position in word: voiceless lenis word-initially, e.g. /b/ → [b̥] / #__ , and voiced word-internally, e.g. /b/ → [b] / V_V.

Register also correlates to a certain extent with overall pitch height: upper register tones have pitch contours mostly in the upper half of the pitch range and lower register tones have pitch contours mostly in the lower half of the pitch range. There is, however, considerable overlap between the high and low register tones' F0 values. In order to show this better, and to move from individual values to values representative of the variety, figure 2 is a plot of three speakers' normalised tones (z-score normalised F0 plotted against normalised duration [19 20]. The individuals' normalised F0 trajectories cluster fairly tightly except for the mid and low rise tones, for which there seems to be greater between-speaker variation: the female's trajectories rise immediately after onset, whereas the males have a delayed rise (so perhaps they should be kept separate). It can be seen that two thirds of the contour of the upper register mid-rise tone, and half the values of the high fall tone, lie below the mid-range value of 0; and the peak values of the lower register low rise-fall are above the mid-range value. This means one cannot define Register – at least as far as these varieties are concerned – in terms of location of pitch/F0 in the upper or lower half of the pitch/F0 range, as is commonly assumed [2].

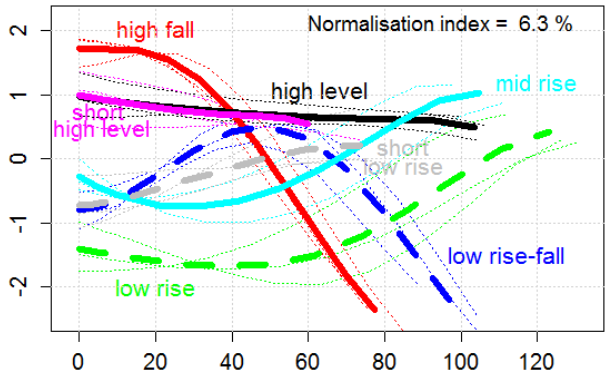


Figure 2: *Normalised values for the seven tones of three Jinshan speakers. Thick lines = mean normalised values. Thin lines = normalised values of individual speakers. Dashed lines = lower register tones.*

Figure 2 also shows how the lower register tones (low rise-fall, low rise, and short low rise) have the same F0 contours as the upper register high fall, mid rise and short high tones respectively, but with a depressed onset. The high level tone lacks a depressed counterpart. It will be shown below how this depression effect can be modeled.

The seven Jinshan tones are thus actually a constellation of pitch, phonation type, duration and segmental quality, and this is again typical of most conservative Wu varieties.

Some of these features are exemplified acoustically in Figure 3, with synchronous wide-band spectrograms and F0, using data from the female speaker. The top panel of figure 3 shows typical segmental and phonatory differences between the upper register high fall tone and lower register low rise-fall tone in words with bilabial stop Onset: [pɪã 51] 冰 *ice* and [b̥ɪ̰ã̰ 231] 平 *flat*. The upper register word has a voiceless unaspirated [p] as allophone of /p/ and has an expected very short VOT lag of less than 1 centisecond. The bilabial Onset in the lower register word is voiceless unaspirated lenis [b̥], which is the word-initial allophone of /b/. A longer duration of about 2 centiseconds between stop release and onset of phonation can be seen. This small difference in VOT, documented for several Wu dialects, presumably reflects a slightly greater distance between the arytenoidal vocal processes at stop release for [b̥] than [p], which in turn reflects the phonatory difference associated with upper and lower register.

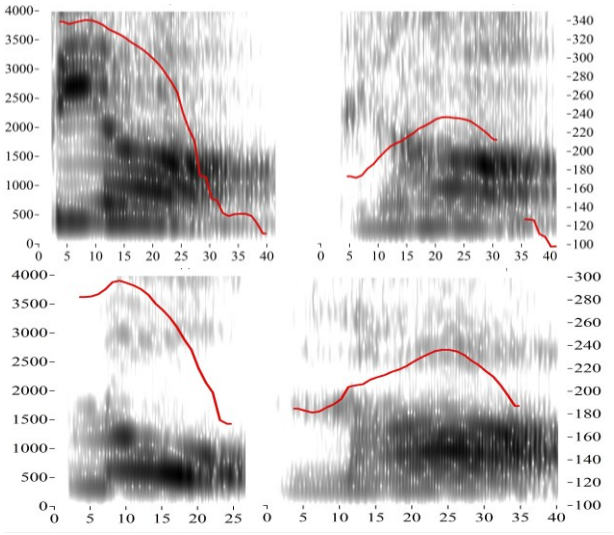


Figure 3: *Acoustics of phonation type and segmental structure differences associated with register differences. Top* = [pɪã 51] 冰 ice *(left) and* [b̥ɪ̰ã̰ 231] 平 flat. *Bottom* = [lɔ 51] 捞 carry *(left) and* [lɥ̩ 231] 狼 wolf. *X-axis = duration (csec.), y-axis left = spectral frequency (Hz), y-axis right = F0 (Hz).*

A far more salient difference between the upper and lower register words is the phonation itself: modal in upper register and breathy in lower. In the lower register word [b̥ɪ̰ã̰ 231] about the first third of the Rhyme – lasting to about the peak F0 point at csec. 20 – shows noisy periodicity, especially clear in the noise-excited F2. This is absent from the modally phonated upper register word (the abrupt spectral change at ca. csec.12 reflects the opening of the velum for the nasalisation in /ɪã/.)

It has often been assumed, e.g. [21 p.91, 22], that breathy voice in Wu is a function of syllable onset *stops*; but in Jinshan, and many other Wu varieties, it characterises all low register words irrespective of whether the Onset is stop, fricative, sonorant, glide or zero. As illustration, the bottom panel of figure 3 shows typical segmental and phonatory differences between the upper register high fall tone and lower register low rise-fall tone in words with *sonorant* Onset: [lɔ 51] 捞 *to carry* and [lɥ̩ 231] 狼 *wolf*. Once again, in the low register word, noisy periodicity is evident in the higher frequency regions (F3, F4) over the first 10 centiseconds after Rhyme onset, as well as during the /l/. Attenuated broadband energy extending from F2 downwards is also seen after Rhyme onset.

3.3. Quantification of phonation type differences

In order to quantify the phonation type differences associated with register, *VoiceSauce* was used to extract spectral slope measures expected to correlate with breathy *vs* modal phonation type. Common practice is to sample parameters at a single point in the tone's time course. It is more informative, however, to quantify the whole of the time course of the parameter. The methodology is described in [23].

Figure 4 shows the time course of two common phonation type parameters: the difference in amplitude between the fundamental and the second harmonic; and the difference in amplitude between the fundamental and the harmonic closest to the first formant centre frequency. *VoiceSauce* calls these "H1H2c" and "H1A1c", where *c* stands for corrected for vowel quality (i.e. using an all-pole LPC transfer function). To provide even tighter control, tokens with non-high vowel nuclei and non-nasal Onsets were used. This means that the estimation of the energy of the fundamental will not be compromised by its being in the vicinity of a low F1 associated with a high vowel or, with nasal Onsets, the lowest nasal formant.

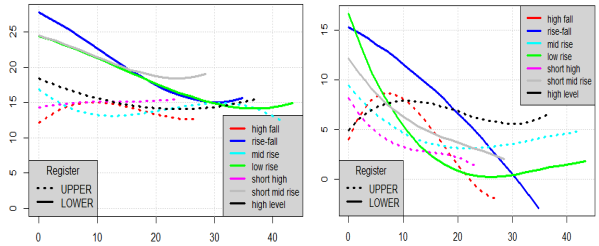


Figure 4: *Time course for VoiceSauce parameters H1A1c (left) and H1H2c for the seven tones of a female Jinshan speaker showing differences between high and low register tones. X-axis = duration (csec.); Y-axis = VoiceSauce parameter (dB).*

Figure 4 shows a clear difference in both parameters associated with register: low register tones have higher values at Rhyme onset than lower. Thus immediately after Rhyme onset the lower register tones have considerably more low frequency energy. This difference disappears towards the end of the Rhyme – more quickly for H1H2c than for H1A1c.

Now, H1A1c and H1H2c differences can be found intrinsically varying with F0 in tones without extrinsic phonation type, e.g. Cantonese [23]. The crucial finding for Jinshan (and other Wu dialects) is that the phonatory parameter is independent of F0. This can be seen by comparing the phonatory parameters of tones *of different register but similar F0*. Fortunately, Jinshan allows us to do this with the lower register rise-fall tone (plotted in blue) and the upper register mid rise tone (plotted in cyan). Figures 1 and 2 show that both these tones have similar F0 over the first 10 centiseconds of their Rhyme. Figure 4 shows their phonatory parameters are very different over this stretch, however: the low rise-fall tone has much greater lower frequency energy. This indicates that, rather than constituting an intrinsic accompaniment to a deliberate low pitch onset in

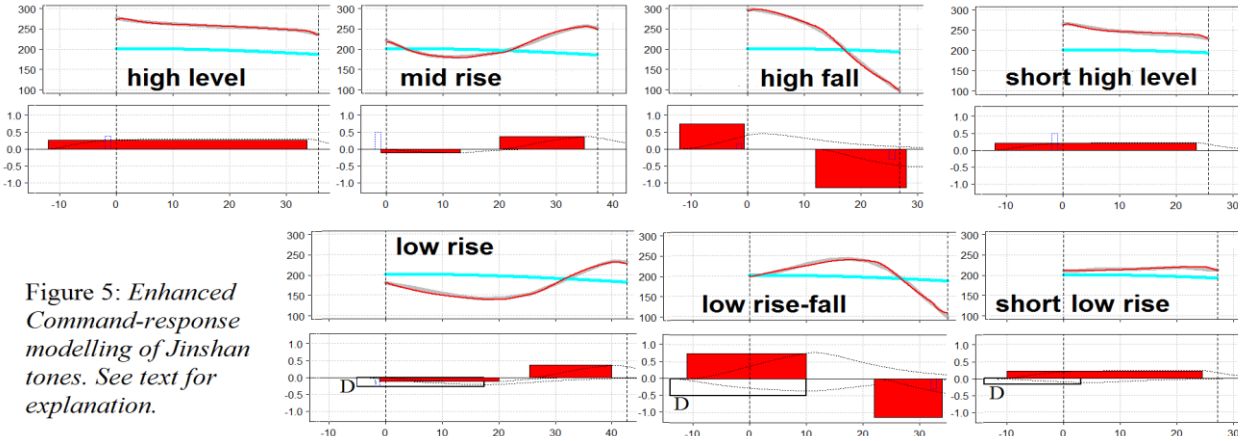


Figure 5: *Enhanced Command-response modelling of Jinshan tones. See text for explanation.*

the low register tones, it is actually the other way round. A low pitch onset is not the cause of the breathy phonation; it is the deliberate breathy voice phonatory setting which causes the low pitch onset. This in turn allows a tonological interpretation of the low register tones as having underlyingly *the same tonal target* as the high register tones but with an additional component which results in depression and breathy voice. The best guess as to the articulatory nature of this component is a constricted epilarynx, an insight of the *larynx-as-articulator* concept described in [24]. The following section shows how this can be modeled phonetically.

3.4. Modeling interaction of phonation type and tone

The interaction between phonation type and tone was quantified using Fujisaki's *command-response* (CR) model for tonal F0 [13 - 15 ch.3]. This model, which has been applied to several tone languages including Mandarin, Cantonese and Shanghai [25 - 27], factors the time-varying F0 into two types of component, both modelled as impulses of given amplitude and duration. A *tonal* component represents the response of the speech production mechanism – in this case the *pars recta* of the crico-thyroid – to impulse commands for implementing tone. The second, or *phrasal*, component represents a much slower time-varying response and accounts for more gradual, declinational, change in F0 throughout an utterance or intonational phrase.

As it stands, the CR model does not allow for anything other than the phrasal and tonal commands, and the tonal commands must not overlap. In order to model the effect of the phonation type in lower register tones, and its interaction with tone, we significantly modified the model to incorporate an additional component which, because it is produced by a different mechanism, can overlap with tonal commands. We have called this a depression component because it models the effect of lowering the F0 at the onset of the Rhyme.

Figure 5 shows the CR modelling of the seven Jinshan tones of the female speaker. Each tone is represented vertically in two panels. The bottom panel shows the impulses and impulse responses: tonal impulses are shaded in red; depression impulses are unshaded and marked with a "D"; impulse responses are plotted in a thin line. The top panel shows, in red, the F0 predicted from the summation of the impulse responses. The speaker's actual tonal F0 is plotted in grey, but is difficult to see as the fit between predicted and observed F0 is very good: mean squared errors range from 3.5 Hz for low rise-fall, down to 1.5 Hz for high level. The cyan line represents the response of the system with just the phrasal command.

It can be seen that the speaker's high level tone is modelled with a single tone impulse, its slight F0 decay accounted for by the decay in the phrasal command. The short high tone is similar, but without such a long tonal command. The high fall tone is modelled with an early positive and a later negative tonal command. The mid rise tone has the opposite arrangement, with an early slightly negative and a late positive tonal command. These results are typical.

It can be seen in figure 5 that each of the lower register tones has been generated *with tonal impulses of the same amplitude as their high register counterparts* plus an additional depression impulse. (The relative *timing* of the impulses has to be such to accommodate the durational differences between the upper and lower register tones, and is not exactly the same.)

This CR analysis-by synthesis buttresses a tonological analysis of the seven tones which allows upper and lower register pairs of tones to be represented with the *same* tonal target, whilst differing in a depression component. Thus the high fall and low rise-fall tones can be analysed as sharing a falling [HL] tonal target, but with the low rise-fall tone having in addition a depression component: [HL, D]. Moreover, the CR analysis supplies an obvious productional interpretation of the tonological constructs H and L as crico-thyroid and strap muscle activity, with D as epilaryngeal constriction. The same analysis applies *mutatis mutandis* for the other tones. Mid rise [MH] and low rise [MH, D] tones share a rising [MH] tonal target, and short high [H] and short low rise [H, D] tones both have a [H] tonal target. The fact that the lower register tones can be shown to have the same tonal target as their upper register counterparts also helps explain the fact that they pattern together in Jinshan's typically complex tone sandhi.

4. Summary

This paper has described the seven isolation tones of Jinshan, a dialect where tone is not just pitch but a constellation of pitch, phonation type, duration and segmental effects. Quantification of the phonation type differences associated with register using conventional spectral slope parameters showed that they were extrinsic and not a function of F0. An extended version of the CR model was then used to demonstrate how the lower register tones can be modeled with the same tonal commands as the upper register tones, but with the addition of a depression component.

5. Acknowledgements

Many thanks to all our Jinshan informants, but especially to teacher 卢迅, for taking their time to read out such a long list so carefully for us. Thanks also to our three anonymous reviewers for some extremely useful comments: we have restructured the paper to take many of these into account.

6. References

[1] Henderson, E., "The topography of certain phonetic and morphological characteristics of South East Asian languages", *Lingua* 15: 400-434, 1965.
[2] Yip, M., Tone, CUP, 2002.
[3] Hyman, L., "Word Prosodic Typology", Phonology 23: 225-257, 2006.
[4] Chao Y. 趙元任，現代吳語的研究 Studies in the Modern Wu Dialects, Tsing Hua College Research Institute Monograph 4, 1928.
[5] Qian N. 钱乃荣，当代吴语研究 [Studies in the Contemporary Wu Dialects], Shanghai Educational Press, 1992.
[6] Zhang J., A Sociophonetic Study on Tonal Variation of the Wúxī and Shànghǎi Dialects, LOT Netherlands Graduate School of Linguistics, 2014.
[7] Gao, J-Y. and Hallé, P., "Are Young Male Speakers Losing Tone 3 Breathyness in Shanghai Chinese? An Acoustic and Electroglottographic Study." Proc. 2nd International Congress on the Phonetics of the Languages in China, 163-166, 2013.
[8] Editorial office for linguistics teaching and research, Hanyu Fangyan Cihui 漢語方言詞彙 [Chinese Dialect Vocabulary], 文字改革出版, 1964.
[9] http://philjohnrose.net/Wu_tones/index.html
[10] Rose, P., "Complexities of Tonal Realisation in a Right-Dominant Chinese Wu dialect – Disyllabic Tone Sandhi in a Speaker from Wencheng", Journal of the South East Asian Linguistics Society 9: 48-80, 2016.
[11] Shen R. and Rose, P., "Preservation of Tone in Right-Dominant Tone Sandhi: A Fragment of Disyllabic Tone Sandhi in Maodian Wu Chinese", in C. Carignan & M. Tyler [Eds] *Proc. 16th Australasian Int'l Conf. on Speech Science & Technology*: 345-348, Sydney, 2016.
[12] Shue, Y.L., Keating, P., Vicenik C. and Yu, K., "VoiceSauce: A Program for voice analysis", in *Proc. 17th Int'l Congress of Phonetic Sciences*, Hong Kong: 1846-1849, 2009.
[13] Fujisaki, H., "Dynamic Aspects of Voice Fundamental frequency in Speech and Singing", in F. MacNeilage [Ed], The Production of Speech, Springer: 39-55, 1983.
[14] Fujisaki, H., "In Search of Models in Speech Communication Research." Proc. *INTERSPEECH,* Brisbane, Australia: 1-10, 2008.
[15] Mixdorff, H., *Intonation Patterns of German – Model-based Quantitative Analysis and Synthesis of F0 Contours*, Ph.D. TU Dresden, 1998.
[16] Rose, P**.,** "Tonation in Three Chinese Wu Dialects"**,** *Proc. Int'l Congress of Phonetic Sciences*, (no page numbers), Glasgow, 2015.
[17] Rycroft, D., "Tone in Zulu Nouns", African Language Studies 4: 43-68, 1963.
[18] Rose, P., "Independent depressor and register effects in Wu dialect tonology: Evidence from Wenzhou tone sandhi", *Journal of Chinese Linguistics* 30(1): 39-81, 2002
[19] Rose, P., "Considerations in the normalisation of the fundamental frequency of linguistic tone", Speech Communication 6(4): 343-352, 1987.
[20] Rose, P., "Comparing Normalisation Strategies for Citation Tone F0 in Four Chinese Dialects", in C. Carignan & M. D. Tyler [Eds], *Proceedings 16th Australasian Int'l Conf. on Speech Science & Technology*, Sydney: 221-224, 2016.
[21] Ramsey, S., The Languages of China, Princeton University Press, 1987.
[22] Cao J. and Maddieson, I., "An exploration of phonation types in Wu dialects of Chinese", Journal of Phonetics 20: 77-92, 1992.
[23] Rose, P., "Variation in Spectral Slope and Interharmonic Noise in Cantonese Tones", Proc. *INTERSPEECH,* Shanghai, 2020.
[24] Esling, J., Moisik, S., Benner, A. and Crevier-Buchman, L., Voice Quality – The Laryngeal Articulator Model, CUP, 2019.
[25] Fujisaki, H., Hirose, K., Hallé, P. and Lei H., "Analysis and modeling of tonal features in polysyllabic words and sentences of the Standard Chinese", Proc. *ICSLP 1990, Kobe, Japan:* 841-844, 1990.
[26] Fujisaki, H., Ohno, S. and Gu W., "Physiological and Physical Mechanisms for Fundamental Frequency Control in Some Tone Languages and a Command Response Model for generation of Their F0 Contours", *International Symposium on Tonal Aspects of Languages*, Beijing, 2004.
[27] Gu W., Hirose, K. and Fujisaki, H., "Analysis of Shanghainese F0 Contours based on the Command-Response Model", Proc. ICSLP: 81-84, 2004.

Just listen: Describing phonetic variation in the word *just*

Ben Gibb-Reid[1], Paul Foulkes[1], Vincent Hughes[1] & Traci Walker[2]

Department of Language and Linguistic Science, University of York, UK[1],
Division of Human Communication Sciences, University of Sheffield, UK[2]
{ben.gibb-reid|paul.foulkes|vincent.hughes}@york.ac.uk, traci.walker@sheffield.ac.uk

Abstract

This study analyses the phonetic variation of the word *just* according to its pragmatic function and surrounding context. Analysis was made of the realisation of its four canonical segments (in Standard Southern British English or SSBE these are /d͡ʒ/, /ʌ/, /s/ and /t/) alongside duration, centre of gravity and formant measures. It was found that tokens of *just* with a discourse function are more likely to exhibit phonetic reduction than adverbial tokens. For instance, discourse *just* has shorter and centralised vowels and a higher rate of vowel and /t/ elision. This suggests that speakers signal different functions of *just* via segmental realisation. Analysing *just* in phonetic detail within its pragmatic and contextual environment describes how the word is shaped in its representation. Understanding the phonetic detail of words helps describe their patterns of social variation.
Index Terms: discourse-pragmatic variation, acoustic phonetics, segmental phonology

1. Introduction

Discourse-pragmatic variables (DPVs) are complex polyfunctional linguistic items which serve to "express speaker stance; to guide utterance interpretation and to structure discourse" [15, p.4]. They include markers (*like, yeah, just*), phrases (*you know, I mean*), interjections (*ah, oh*) and longer strings. Their structural and interactional patterning has been analysed by referring to frequency of use, functions, grammaticalization processes, and social variation. The properties of DPVs also vary according to function and prosodic context, as has been shown with *like* [10,11,18]. However, it is rare for these discourse features to be analysed phonetically, even though word-level variation is important in understanding social patterns of production [12]. *Just* has been studied with a focus on its various discourse functions and overall frequency of occurrence. Adverbial functions of *just* ('to be precise' and 'to reference the recent past') are pragmatically distinct from emphatic (restricting / intensifying meaning) or planning (when used as a 'filler') functions [1,2,4]. Phonetic variation of *just* has not been the subject of detailed study, with the exception of [8]. The present study compares phonetic variation across pragmatic functions and segmental contexts of *just*. The study thus contributes to a growing body of work which analyses phonological patterns via individual words in context [12,16]. As assumed by exemplar theory, words stand at the intersection between linguistic and social meaning, and analysis of their phonological and contextual variation gives insight into individual speaker and language change [12].

2. Background

Just is a highly frequent word in spoken English. In the 2014 spoken British National corpus, it is the 27th most frequent word at 0.75 occurrences per 100 words (up from 42nd in the 1994 edition) [13]. *Just* is also increasing over apparent time in Toronto English [19]. These frequency measurements, however, combine multiple functions of *just*.

Other studies show differences in how the various functions are distributed socially and stylistically. Aijmer [1,2] highlights the use of *just* as a metalinguistic marker which occurs between utterances and marks them as expressing attitudes or emotions. Beeching [4] categorizes minimizing, intensifying and filler functions in spoken British English and compares use across ages and sexes. Woolford [22] analyses *just* functions quantitatively in pre-verbal contexts across speaker age and sex in Tyneside English. Across all quantitative studies, *just* is used more by younger speakers, and there is some indication in the BNC that female speakers use it more than males [13].

The word *like* has been analysed with reference to its function and pronunciation in a number of studies. In New Zealand, London and Edinburgh, when *like* is utilized as a quotative marker (e.g. "I was *like*, 'sure you can borrow that'"), it tends to have a more monophthongal vowel and a shorter /l/ to vowel ratio than when it is a verb, conjunction or discourse marker (e.g. "I *like* cheese" / "I feel *like* rubbish" / "You don't, *like,* even know"). Discourse marker and quotative *like* also had a more reduced /k/ [11,18]. Schleef and Turton [18] argue that vowel quality in *like* is linked to its position near intonation boundaries, such that a pre-pausal *like* tends to have a diphthongal vowel, whereas *like* in continuing speech tends to be more monophthongal. Quotative *like* is also more monophthongal compared to other functions, but this is caused by its typical occurrence surrounded by speech. *Like* also varies in pronunciation according to speaker stance and style [10]. More generally, highly frequent words are more likely to be phonetically reduced [6]. As *just* occurs frequently, it can therefore be expected to reduce, for example with vowel and /t/ elision and vowel centralisation. Vowels have a more central F2 when occurring in highly frequent words, and "less frequent words [are] less apt to undergo lenition, since they are more in need of the extra phonetic clarity afforded by distinct, non-lenited articulation" [9, p.103]. Lenition of *just* might also be expected to follow general patterns relating to /t,d/ in coda obstruent clusters. In a recent study of 14,000 tokens in British English, pre-consonantal /t,d/ had an overall deletion rate of 75% [3]. /t,d/ were also more likely to be deleted in frequent words, irrespective of the following phoneme, and *just* specifically had a rate of /t/ deletion over 50% for almost all speakers. Generally, research on /t,d/ deletion has found that there is a spectrum of following contexts from more to less elision: obstruents > liquids > glides > following vowels and pauses [20]. By analysing the variation in *just* across context and pragmatic meaning, the present study aims to understand phonological changes at broader levels than the segment, and to align form with function.

2.1. Predictions

Just is predicted to vary phonetically in several ways: (i) *Just* will be frequent and exhibit high rates of vowel centralization and /t/ elision. (ii) Functions of *just* will differ from one another in vowel quality and segment reduction, similar to *like*. (iii) Pre-pausal *just* will show less phonetic reduction than *just* in continuing speech.

3. Methodology

3.1. Data

The data for this study is taken from the DyViS corpus, task 1 [14]: 100 18-25 year old male speakers of Southern Standard British English recorded in simulated police interviews. Participants were asked a set of questions by a researcher acting in the role of a police officer and answered spontaneously, guided by maps and pictures. Each recording lasts 20-30 minutes. Sound files were listened to, segmented, transcribed and measurements were extracted using Praat (6.2.12) [5]. R Studio [17] and tidyverse [21] was used to process the results, run statistical tests, manipulate, tidy and visualise the data. To interpret the results, various linear mixed effects models were run, utilizing the lme4() package [7]. For categorical data like vowel and /t/ elision this took the form of regression models. For continuous data such as vowel formant readings, ANOVA model comparisons were run initially to identify which variables contributed significantly to predicting the data. Then, the best model was run including only the significant predictors.

3.2. Features and coding

Each token of *just* was extracted for analysis and coded according to a range of variables. *Just* functions were based on [22]. The categories are shown in Table 1, (1)-(6). Examples indicate interviewers 'I' and participants 'P'.

An auditory analysis was undertaken of every token to identify the presence/absence of each of the four canonical segments, and their durations were also measured. The hold phase of the initial affricate /d͡ʒ/ was not included in duration measurements because the onset of the hold was not identifiable if preceded by silence. /d͡ʒ/ was measured from the release of the plosive (indicated by a burst of energy and an aperiodic waveform) to the beginning of open approximation and voicing for the vowel (indicated by clear formant structures, periodicity and a higher amplitude). The end of the vowel was identified where the waveform became aperiodic, reflecting /s/ frication, and the latter boundary for /s/ was defined by either closure for /t/ (indicated by a drop in amplitude and a lack of high-frequency activity in the spectrogram), or the beginning of an immediately following sound (for example, another vowel - "just after"). Where the boundaries between sounds were unclear, they were marked as being unfit for duration or formant measurements and no values were taken. Tokens of KIT, STRUT and the vowel from the filled pause *um* were also extracted from speakers for comparisons of vowel quality. STRUT was chosen as this is the citation form lexical set for *just* and KIT was utilized as others [8] have found that *just* vowels pattern closely with it. *Um* was analysed as a typically more centralized vowel for comparison with the degree of *just* vowel reduction. Following contexts were categorised as either pauses (over 100ms of silence), consonants or vowels.

Category	Sub-category	Definition	Example
Adverb	Particularizer	to do with location, meaning 'precisely/exactly'	(1) I: What do you see from the window? P: There's a tour bus leaves, a city tour bus leaves from **just** down the road.
	Temporal adverb	referring to the recent past or simple perfect	(2) P: he **just** started working there.
Restrictive		meaning 'nothing other than', diminutive	(3) I: Did you give someone a lift after work? P: Um, no it was **just** me driving home.
Discourse	Evaluative	meaning 'no more than' or 'merely / simply one of a few'	(4) I: Do you know someone who works there? P: Not more than **just** having them serve me a drink.
	Intensifier	boosts or maximizes the force of the focused item, meaning 'really'	(5) I: you didn't hang out together. P: I **just** don't know him I'm afraid
Filler		when just is repeated or there is some cut-off and the intended meaning is unclear	(6) P: I tend to park my car **just** be- just behind the hairdresser's.

Table 1: Just *function categories based on [22], aside from filler which is taken from [4].*

	Consonant		Pause		Vowel		
	%	N	%	N	%	N	TOTAL N
Discourse	69.4	347	16.6	83	13.2	66	500
Adverbial	48.3	262	3.5	19	48.2	261	542
Restrictive	86.3	120	2.9	4	10.8	15	139
Filler	27.4	26	35.8	34	31.6	30	95
TOTAL	59.2	755	11.0	140	29.2	372	1276

Table 2: *Distribution of* just *functions according to following context.*

4. Results

The spread of *just* across its functions and following contexts is shown in Table 2. In total, 1,276 tokens of *just* were extracted from the dataset. On average this equates to 0.88 occurrences per 100 words across the corpus. Discourse and adverbial functions of *just* are the most common, followed by restrictive and filler tokens. Tokens which function as fillers are the most likely to occur before a pause. Discourse and restrictive *just*, however, are more likely to occur before a consonant (the most likely place for /t/ elision) than a vowel or pause. Adverbial *just* occurs equally often before consonants and vowels.

1,019 vowel tokens were deemed suitable for formant analysis. The spread of midpoint F1 and F2 measurements are displayed in Figure 1. *Just* vowels generally had F1 and F2 values nearer to KIT and the vowel in *um*, and are distinct from STRUT (a pattern reported elsewhere, e.g. [8]). They are central, with mean F1 at 395 Hz and mean F2 at 1544 Hz.

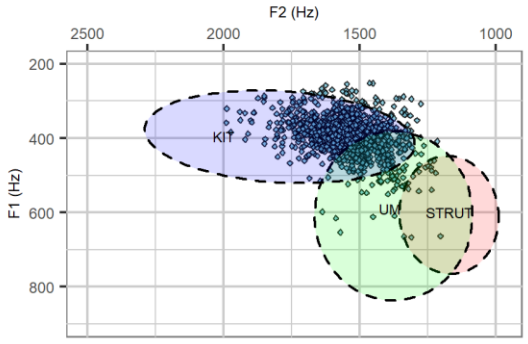


Figure 1: *F1-F2 plot of all* just *vowel midpoints extracted for analysis. Standard deviations of KIT, STRUT and the vowel from* um *shown in ellipses.*

4.1. Variation by function

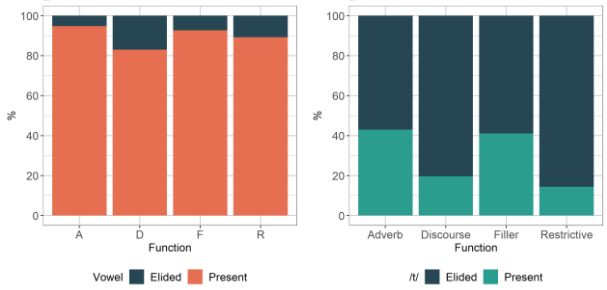


Figure 2: Just *functions by proportion of vowel elision (left) and /t/ elision (right). A=adverb, D=discourse, F=filler, R=restrictive.*

. The spread of vowel elision across functions is illustrated in Figure 2 (left panel). Overall, vowel elision occurs in 10.5% of tokens, and /t/ elision in 65.3%. Across functions, discourse *just* had the highest degree of vowel elision (16.8%). This was significantly greater than for adverbial (z(0.23)=5.0, p=<0.001) and restrictive (z(0.3)=-2.0, p=<0.05) tokens. Restrictive *just* had the next highest rate of vowel elision at 10.8%, though this was not significantly higher than adverb or filler tokens.

The proportion of /t/ elision across functions is also shown in Figure 2 (right panel). Tokens of *just* with a restrictive function are most likely to have elided /t/ (85.6%), followed by discourse *just* (80.0%). Table 2 shows that restrictive and discourse tokens typically occur pre-consonantally, which indicates that /t/ elision is linked to segmental context. This is explored in section 4.2. Again, overall we see discourse *just* showing some phonetic reduction, though this was not significantly different from adverbs or restrictive tokens.

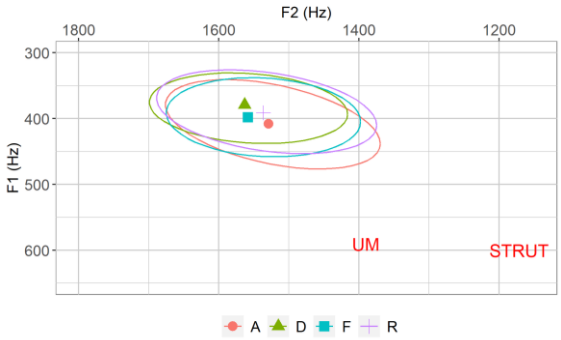


Figure 3 : *F1-F2 plot of all* just *vowel midpoints across functions. Ellipses indicate the standard deviation of values per function. A=adverb, D=discourse, F=filler, R=restrictive.*

F1 and F2 values for *just* are plotted by function in Figure 3. Of the four functions, the one closest to STRUT is adverb. Discourse tokens were furthest removed from the STRUT mean. Adverb tokens had overall higher F1 values (mean 408 Hz) than discourse tokens (mean 379 Hz); this was a significant difference (t(3.4)=-7.6, p=<0.001). For F2, adverbial tokens had slightly lower values (1529 Hz) than discourse tokens (1563 Hz) though function was not a significant predictor in model comparisons for F2 values. Discourse *just* tokens also had significantly lower F1 values than restrictive tokens (t(4.7)=-3.5, p=<0.001). This suggests that discourse *just* has closer vowels than adverbial or restrictive *just*.

In terms of segment duration, there were no clear patterns or significant effects for /d͡ʒ/. For vowels, restrictive and discourse tokens exhibited slightly shorter durations overall (means 52 and 53 ms) than adverb tokens (56 ms) but function did not significantly predict vowel duration (x^2(3)=4.0, p=0.27). Discourse tokens had a longer /s/ (mean 78 ms) and /t/ (mean 65 ms) whereas adverbial tokens had shorter durations (/s/= 68 ms, /t/=54 ms). Both /s/ (t(<0.1)=-4.4, p=<0.001) and /t/ (t(<0.01)=-5.8, p=<0.001) were significantly longer when *just* had a discourse rather than adverbial function. Filler *just* had the longest /s/ (t(<0.1)=-4.3, p=<0.001) and /t/ (t(<0.1)=-4.5, p=<0.001). However, discourse and filler tokens also exhibited large standard deviation for /s/ (discourse: 33 ms, filler: 59 ms) and /t/ durations (discourse: 36 ms, filler: 37 ms).

4.2. Variation by following contexts

Vowel elision rates for *just* across following contexts are displayed in Figure 4 (left panel). *Just* preceding a vowel had the least vowel elision (6.2%), with higher rates when occurring before a consonant or a pause (12.7% and 9.3% respectively). None of these differences between following contexts and vowel elision were significant.

The segments following *just*, however, affected the rate of /t/ elision to a greater degree, as shown in Figure 4 (right panel). /t/ elision had a similar pattern to vowel elision. Pre-vocalic tokens were the least likely to have /t/ elided (22% elision). This was significantly lower than for pre-pausal tokens (z(0.3)=8.5, p=<0.001). Pre-pausal *just*, in turn, had less /t/ elision (72%) than pre-consonantal *just* (92%) and this was also significant (z(0.3)=-5.9, p=<0.001).

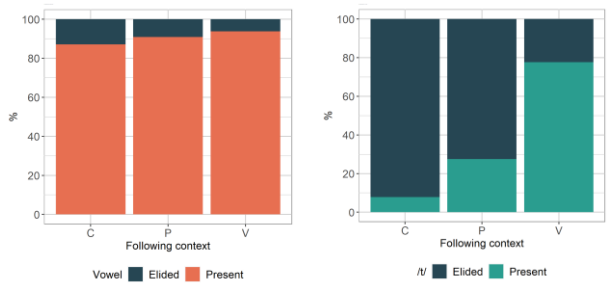


Figure 4: Just *following contexts by proportion of vowel elision (left) and /t/ elision (right). C=consonant, P=pause, V=vowel.*

It is clear from Table 2 that some functions occur in certain contexts more than others. Specifically, restrictive and discourse tokens occur more pre-consonantally than adverb tokens do. An interaction of function and following context was therefore included in a /t/ elision regression model with a smaller dataset (one without filler or pre-pausal tokens). This yielded a significant result when comparing discourse with adverb across pre-consonantal and pre-vocalic contexts (z(0.5)=2.15, p =0.031). This suggests that *just* /t/ elision is determined more by the following context than its function. A similar model for vowel elision did not yield any significant interactions.

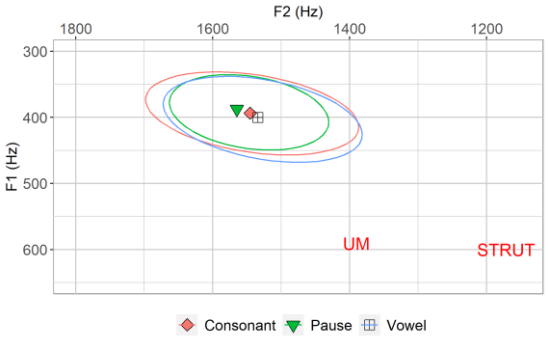


Figure 5: *F1-F2 plot of all* just *vowel midpoints across following contexts. Ellipses indicate the standard deviation of values per function.*

Finally, Figure 5 plots formant values across following contexts. Pre-vocalic and pre-consonantal tokens were close in mean vowel quality, with pre-vocalic tokens having a slightly higher F1 (400 Hz versus 393 Hz respectively) and a lower F2 (1534 Hz versus 1545 Hz). The distinct category here is pre-pausal. When *just* occurred before a pause it had a higher F2 (mean 1564 Hz) and a slightly lower F1 (mean 388 Hz) than when it occurred before consonants or vowels. This suggests that pre-pausal *just* tends to have more centralized vowels. However, in model comparisons, following contexts were not significant predictors for F1 ($X^2(2)=0.11$, $p=0.94$) or F2 ($X^2(2)=0.39$, $p=0.82$).

For following contexts, the overall pattern of segment duration points towards pre-consonantal just being shorter. With a mean of 166 ms it was far shorter than pre-pause and pre-vocalic (both are mean 224 ms) just tokens. As with function, this was starkest across /s/ and /t/ durations. For /s/, pre-consonantal tokens had a mean length of 66ms, significantly lower than pre-vocalic tokens at 78 ms (t(<0.1)=8.1, p<0.001) and pre-pausal tokens at 97 ms (t(<0.1)=6.6, p<0.001). For /t/, pre-consonantal tokens had a mean length of 60.2 ms, almost identical to pre-vocalic tokens which had a mean of 59.7 ms, but significantly shorter than pre-consonantal tokens at 105 ms (t(<0.1)=7.5, p<0.001). The pattern is less clear for /d͡ʒ/ and vowel durations with only a few milliseconds between the mean values of each following context

5. Discussion

Overall, *just* is a highly frequent word in the data. At 0.88 tokens per 100 words, it occurs more often than in the 2014 British National Corpus [13] or Tagliamonte's Toronto corpora [19]. This discrepancy may have something to do with the map task participants were asked to undertake – by describing map features, it is possible that there was a higher proportion of particularizer *just* (e.g. "the barber shop is just down the road from…").

Though the vowel in *just* is historically STRUT for SSBE, in the present data it is in the direction of KIT but does not fit either category (see Figure 1). This underlines the need to consider word-specific phonetics, as words pattern uniquely and are individual units. The high frequency of *just* is also a potential contributing factor to its vowel production.

Speakers indicate different pragmatic functions of *just* by their choice of segments – aligning phonetic resources with pragmatic meaning. Discourse *just* was more likely to exhibit longer /s/ and /t/, and to have more vowel elision and lower F1 values than other function categories. Restrictive *just* had shorter /s/ and /t/ durations, higher rates of /t/ elision and higher F1 values. Filler *just* exhibited the longest /s/ and /t/ durations. Adverbial *just* exhibited shorter /s/ and /t/ durations, less vowel elision and the highest F1 values. The following context of *just* predicts the rate of /t/ elision with a hierarchy of more to less elision: consonants > pauses > vowels, similar to the pattern found in other studies [20]. Although segment duration and vowel elision and quality are predicted by *just* pragmatic function, following context is a better predictor of /t/ elision. This corroborates findings on *like* [11,18], where there is an interaction between surrounding contexts and token functions.

In terms of phonetic reduction, discourse *just* had higher rates of vowel and /t/ elision and a more central vowel. This is beyond a simple frequency effect as the most frequent 'word' here was actually adverbial *just* (42% of tokens), followed by discourse *just* (39%). Pre-consonantal *just* was the most reduced context, although it patterned closely with pre-pausal *just* in vowel elision and vowel quality. Future work should consider specific following/preceding segments, allowing an exploration into coarticulation effects.

By describing the phonetic variation of *just* with detailed acoustic analyses, it is shown that speakers utilize phonetic resources to indicate social and pragmatic meaning. This is alongside adherence to phonological patterns of contextual variation. Words can display specific phonetic patterns [12, 16], and understanding their variation and meaning sheds light on how we communicate indexical and pragmatic meanings in grammatical as well as lexical words.

6. References

[1] K. Aijmer, English discourse particles: Evidence from a corpus. John Benjamins Publishing, 2002.

[2] K. Aijmer, "'Just' and multifunctionality," in Contexts - Historical, Social, Linguistic: Studies in Celebration of Toril Swan, K. McCafferty, Bull, Tove & Killie, Kristin Ed.: Peter Lang, 2005, pp. 31-47.

[3] M. T. Baranowski and D. Turton, "TD-deletion in British English: New evidence for the long-lost morphological effect," Language Variation and Change, vol. 32, no. 1, pp. 1-23, 2020.

[4] K. Beeching, Pragmatic Markers in British English: Meaning in Social Interaction. 2016.

[5] P. Boersma and D. Weenink, Praat: doing phonetics by computer. (2022). Accessed: 18 May 2022. [Online]. Available: https://www.praat.org

[6] J. Bybee, "Word frequency and context of use in the lexical diffusion of phonetically conditioned sound change," Language variation and change, vol. 14, no. 3, pp. 261-290, 2002.

[7] B. Bolker, lme4: Mixed-effects models. (2022). Accessed: 18 May 2022. [Online]. Available: https://github.com/lme4/lme4/

[8] CCC, "It's not just a sound change: linking phonetic and pragmatic change in a discourse-pragmatic marker.," (forthcoming).

[9] A. J. Dinkin, "The real effect of word frequency on phonetic variation," University of Pennsylvania Working Papers in Linguistics, vol. 14, no. 1, p. 8, 2008.

[10] K. Drager, "Constructing style: Phonetic variation in discursive functions of like," in Discourse-Pragmatic Variation and Change in English: New Methods and Insights, H. Pichler Ed. Cambridge: Cambridge University Press, ch. 10, pp. 232-251, 2016.

[11] K. Drager, "Sociophonetic variation and the lemma," Journal of Phonetics, vol. 39, no. 4, pp. 694-707, 2011.

[12] J. Hay, "Sociophonetics: The Role of Words, the Role of Context, and the Role of Words in Context," Topics in Cognitive Science, vol. 4, no. 10, pp. 696-706, 2018.

[13] R. Love, C. Dembry, A. Hardie, V. Brezina, and T. McEnery, "The Spoken BNC2014: Designing and building a spoken corpus of everyday conversations," International Journal of Corpus Linguistics, vol. 22, no. 3, pp. 319-344, 2017.

[14] F. Nolan, K. McDougall, G. de Jong, and T. Hudson, "The DyViS database: style-controlled recordings of 100 homogeneous speakers for forensic phonetic research," in International Journal of Speech Language and the Law vol. 16, ed, 2009, pp. 31-57.

[15] H. Pichler, The structure of discourse-pragmatic variation. John Benjamins Publishing, 2013.

[16] J. Pierrehumbert, "Word-specific phonetics" Laboratory phonology, vol 7, no. 1, pp. 101-140, 2006.

[17] RStudio: Integrated Development for R. (2020). R Studio, PBC, Boston, MA, URL http://www.rstudio.com/.

[18] E. Schleef and D. Turton, "Sociophonetic variation of like in British dialects: effects of function, context and predictability," English Language and Linguistics, vol. 22, no. 1, pp. 35-75, 2016, doi: 10.1017/S136067431600023X.

[19] S. Tagliamonte, Teen talk: The language of adolescents. Cambridge University Press, 2016.

[20] S. Tagliamonte and R. Temple, "New perspectives on an ol' variable: (t,d) in British English," (in English), Language Variation and Change, vol. 17, no. 3, pp. 281-302, Oct 2005, doi: http://dx.doi.org/10.1017/S0954394505050118.

[21] H. Wickham et al., "Welcome to the Tidyverse," Journal of Open Source Software, vol. 4, no. 43, p. 1686, 2019, doi: 10.21105/joss.01686.

[22] K. Woolford, "Just in Tyneside English," World Englishes, 2021.

A Gaussian mixture classifier model to differentiate respiratory symptoms using phonated /ɑ:/ sounds

Balamurali B T[1], Hwan Ing Hee[2], Cindy Ming Ying Lin[1], Prachee Priyadarshinee[1], Christopher Johann Clarke[1], Dorien Herremans[1], Jer-Ming Chen[1]

[1]Singapore University of Technology and Design, Singapore
[2]KK Women's and Children's Hospital, Singapore
balamurali_bt@sutd.edu.sg, jerming_chen@sutd.edu.sg

Abstract

An audio-based classification model that differentiates between healthy vs pathological respiratory symptoms using acoustic features extracted from phonated /ɑ:/ sounds is presented. For this, a new dataset of phonated /ɑ:/ sounds, together with a clinician's diagnosis, was compiled and a Gaussian Mixture Model (GMM) using Mel-Frequency Cepstral Coefficients (MFCCs) classifier was used. Despite no significant differences in mean values of the fundamental and formant frequency (F0, F1, F2, and F3) distribution for /ɑ:/ sounds retrieved from healthy vs pathological populations, our /ɑ:/ sound model trained using MFCCs resulted in an accuracy of 81.92% when compared against clinician's diagnosis.

Index Terms: Machine Learning, Phonated /ɑ:/ sound, Respiratory symptoms, Gaussian Mixture Model, Mel Frequency Cepstral Coefficients.

1. Introduction

Many childhood respiratory conditions such as asthma, respiratory tract infections, and allergies, often characterised by the presence of coughing, can present severe challenges to anaesthetists during the perioperative period. Some of these conditions, such as respiratory infections, necessitate deferring surgery. Others, such as asthma, require simply specific anaesthetic care and do not require a postponement. As a result, a correct differential diagnosis of respiratory symptoms is critical to a successful surgery anaesthetic outcome [1, 2]. A majority of paediatric surgical cases are often performed in day surgery, with preoperative screening often performed over the phone by nurses on the day before surgery. However, it is frequently difficult for nurses to identify respiratory conditions only using a history of coughing or other specific articulatory cues provided by children over the phone. This can result in a delayed diagnosis and the cancellation of planned surgery.

In this investigation (as a follow-up study to [3, 4]), we present a unique /ɑ:/ sound dataset and a model that distinguishes between healthy and pathological respiratory symptoms using acoustic features extracted from /ɑ:/ sounds (i.e., /ɑ:/ vowel as in the word 'Father'). Because each respiratory pathology produces its own spectral features owing to the differences in airway dimension, patency, and secretions associated with respiratory pathologies, the articulatory sound may contain cues that can be exploited to classify the underlying respiratory symptoms. The accuracy of the proposed model is validated by comparing the model's output to the clinician's diagnosis.

During clinical examination, physicians frequently instruct patients in general to phonate /ɑ:/ ('aah') to inspect patient's larynx, a procedure that has evolved over the course of many years of customary practice [5–7]. This strategic articulatory gesture gives the patient the ability to lower the middle and rear of the tongue while simultaneously extending the jaw opening, which enables simple visual access to the back of the mouth. From an acoustic standpoint, phonated sounds can indicate physiological changes to the vocal folds, vocal tract, and associated respiratory regions. These changes often include swelling or inflammation of the ear, nose, or throat tissues that are associated with speech, swallowing, or breathing.

Cough sounds or other articulatory sounds are used in research to develop automatic classification models that can differentiate between various respiratory disorders [8–13]. To distinguish between productive and non-productive coughs, Murata used time expanded waveforms paired with spectrograms [14]. Abaza created a setup that uses a combination of airflow parameters and audio parameters of voluntary coughs to detect impaired lung functions [15]. Cough sound analysis has also been used to identify pneumonia more quickly [16]. In this investigation, we present a phonated /ɑ:/ sound dataset and a model trained on the features extracted from this sound that can accurately differentiate healthy and pathological respiratory symptoms.

2. Data Collection

2.1. Subject Recruitment & Audio Recording Procedure

Children from the pathological group (a spectrum of respiratory conditions including asthma, Upper Respiratory Tract Infection (URTI) and Lower Respiratory Tract Infection (LRTI)) were recruited from the Children's Emergency Department, Respiratory Ward, and Respiratory Clinic at KK Women's and Children's Hospital, Singapore. The /ɑ:/ sounds were recorded during the initial presentation at the hospital. Children from the healthy group were recruited from the Children Surgical Unit of at KK Women's and Children's Hospital, Singapore. These /ɑ:/ sounds were recorded at the hospital on the day before surgery (children scheduled for surgery ideally should not have any respiratory infections). A total of 593 /ɑ:/ sounds (typically one to two seconds long duration) were recorded, 467 from children who had respiratory symptoms (pathological) and 126 from children who were healthy (See Table 1 for more details).

A smart phone was used to record the /ɑ:/ sounds at 44.1 kHz from both pathological and healthy children. The recordings were made in a "raw" clinical context, i.e., in-situ with background noise such as conversations, public address announcements, beeping equipment, distant siren sounds. Partic-

Table 1: *Number of instances of /ɑ:/ sounds.*

/ɑ:/	Number of sounds
Healthy	126
Pathological	
• URTI	109
• Asthma	178
• LRTI	180
• Total	467

ipants were requested to phonate /ɑ:/, which were then manually segmented into distinct dataset entries. There are a few cases that have two /ɑ:/ records. Both traditional (fundamental and formant frequency) and automatic audio features (MFCCs) were extracted, where the former were used to investigate if the features (extracted from healthy and pathological group) have equal mean values or not and the latter were used to model Gaussian mixtures. Additionally, the /ɑ:/ sound was further recorded for eight children after they had recovered from respiratory symptoms. This later data was used to conduct a longitudinal study to better understand the evolution of traditional audio features such as F0, F1, F2 and F3.

3. Feature Modelling and Likelihood Ratio

Two distinct Gaussian Mixture Model - Universal Background Models (GMM-UBMs) were used to model features extracted from phonated /ɑ:/ sounds. A Universal Background Model (UBM) was firstly developed in this process using audio feature data pooled across both classes (healthy and pathological), with a Gaussian Mixture serving as the probability density function (optimal fit for this probability density function was found using the Expectation Maximization (EM) algorithm) [17]. This UBM (in this investigation, UBM was created using 256 Gaussian components) is then used to generate a healthy and a pathological model by adjusting the background model to provide a better fit for features extracted from the healthy and pathological articulatory sounds, respectively. Both adaptions are accomplished through the Maximum A Posterior (MAP) technique.

To estimate a likelihood ratio, the conditional probability of the evidence given the hypothesis of whether the articulatory sound belongs to a healthy or pathological subject is evaluated. Likelihood ratio (LR), as the name suggests, is the ratio of two conditional probabilities. In the context of this study, the LR framework gives a quantitative estimate of which group the articulatory sound belongs to:

$$LR = \frac{p(E/H_{Healthy})}{p(E/H_{Pathology})} \quad (1)$$

where $p(E/H_{Healthy})$ computes the conditional probability of E (the evidence) given the hypothesis (H) that articulatory sound is healthy, whereas $p(E/H_{Pathology})$ calculates the probability of evidence given the hypothesis (H) that sound sample is pathologic. The healthy hypothesis is supported by LR values greater than one, whereas the pathology hypothesis is supported by LR values less than one. Values close to one are inconclusive for both hypotheses. From the LR value, the Log-Likelihood-Ratio (LLR) was calculated as $LLR = \log_{10}(LR)$. The sign of the LLR reveals whether the model favors a healthy sound (i.e., positive LLR) or a pathological sound (i.e., negative LLR) and its magnitude reflects how strong that support is [18, 19].

3.1. Features for Modelling GMM-UBMs

GMM-UBMs are modelled using Mel frequency cepstral coefficients (MFCCs) and are extracted as follows. The /ɑ:/ sounds were first divided into frames of 100 ms with 50 ms overlap, after which a hamming window was applied. MFCCs were then extracted from every frame by first calculating the spectrum using discrete Fourier transform. Frequency-related information is extracted by creating a set of overlapping non-linear Mel-filter banks. The logarithm of the energy corresponds to each filter region of the audio spectrum is then estimated. MFCCs are finally derived by taking the discrete cosine transform of this log spectrum [20, 21]. A total of 42 features was extracted per frame, i.e., 14 MFCCs, 14 deltas, and 14 delta-deltas. MFCCs were chosen owing to their effectiveness when they come to audio classification problems [22, 23] whereas deltas and delta-deltas provide valuable cues about audio dynamics and have shown to improve audio classification accuracy [24].

4. Experimental Setup

The model was trained and evaluated using leave-one-out cross-validation, in which the model is trained using all of the data except for one data point, for which a prediction is then made. With 593 data samples, a total of 593 distinct models must be trained; while this is a computationally expensive methodology, it ensures a reliable and unbiased measure of model performance. This computationally expensive methodology, however, limits the use of memory intensive machine learning techniques such as support vector machine (SVM), ensemble learners, and deep neural nets in this investigation.

To assess the performance, Tippett plots and Receiver Operating Characteristics (ROC) were used. Tippett plots illustrate the cumulative proportions of LLR values for both healthy and pathological /ɑ:/ sounds (represented using dotted and solid curves, respectively).The farther apart these curves are, the better the result [25]. In ROC, true positive rates (i.e., sensitivity: the ratio of true positives to the sum of true positives and false negatives) are plotted against false positive rates (i.e., (100—specificity); specificity is the ratio of true negatives to the sum of false positives and true negatives) for various decision thresholds. A perfect model yields a ROC curve that passes towards the upper left corner, indicating greater overall accuracy [26]. This would result in a ROC with an area underneath (AROC) of one. The classification accuracy, sensitivity, specificity, and AROC of the results are also explored to fully comprehend model performance [27].

5. Results

5.1. F0, F1, F2 and F3 Analysis

The hypothesis that the two groups of features (extracted from healthy and pathological group) have equal mean values (null hypothesis) or not was tested using Welch's T-Test. This test was chosen because the two samples have unequal sample sizes and may have unequal variances [28]. This hypothesis testing was done only for the traditional features such as fundamental frequency F0 and formant frequencies (F1, F2 and F3) (extracted using Praat [29]), both of which depend on vocal tract physiology. For e.g., F0 is impacted by the change in mass, and longitudinal tension of the vocal folds. The first formant (F1) is usually influenced by the height of the tongue. The higher the tongue, the lower F1. The position of the tongue from front to back reflects on the changes in the F2 values. When it comes

to front vowels, F2 is often higher than the back vowels. Finer acoustic differences between vowels are made by rounding or not rounding the lips, which mostly affects F2 and F3 [30]. When the vocal folds oscillate during phonation, acoustic energy is not only transmitted 'forwards' to the open lips (resulting in speech sounds), but also 'backwards' to the trachea and lungs. While the lungs are acoustically lossy, some acoustic energy will still back-propagate out through the glottis, contributing to a different vocal quality if the trachea and lungs have changes associated with certain respiratory conditions. The sum total of these physiological changes across the respiratory system will likely influence F0, F1, F2 and F3, in addition to other vocal cues (such as timbre).

A total of two hypothesis tests were carried out. The first compares and contrasts two distinct populations (i.e. on features extracted from healthy and pathological population). The second still compares two populations, but features are extracted when subjects are having respiratory symptoms and then again when they are recovered (i.e., on a longitudinal feature set). The respective results are shown in Tables 2 and 3.

Table 2: *Healthy VS Pathology Welch's T-Test results for F0, F1, F2 and F3.*

Features	p value
F0	0.450
F1	0.070
F2	0.372
F3	0.560

For t-test contrasting the two distinct population of healthy and pathology (Table 2), all the p-values are greater than 0.05 and thus fail to reject the null hypothesis that there is no difference between the mean values of the F0, F1, F2 and F3 extracted from both the population. This result is further verified and illustrated in Figure 1, which uses a box plot and a probability density function to depict the distribution of F0, F1, F2, and F3 extracted from healthy and pathological populations.

Table 3 shows the results of the t-test on the longitudinal feature set. Except for F3, all of the p-values are more than 0.05, indicating that there may be differences in mean values of F3. However, because the number of samples available for this longitudinal analysis is limited (/ɑ:/ from 8 subjects only), this finding should not be generalized (See Figure 2 for distribution).

Table 3: *Longitudinal Welch's T-Test results for F0, F1, F2 and F3.*

Features	p value
F0	0.341
F1	0.241
F2	0.528
F3	0.013

As discussed, there was no significant difference in the mean values of the features extracted during pathology vs. post recovery (except for F3, Table 3), even though the features appeared to evolve in the same direction. When the patients recovered, the values of F0, F1, F2, and F3 increased. This is shown in Figure 3 and this change could be attributed to changes in the participants' vocal cord and vocal tract. However, the trend in F1, F2, and F3 for the same subject is inconsistent (the subject whose F1 drops upon recovery has an increase in F2 or

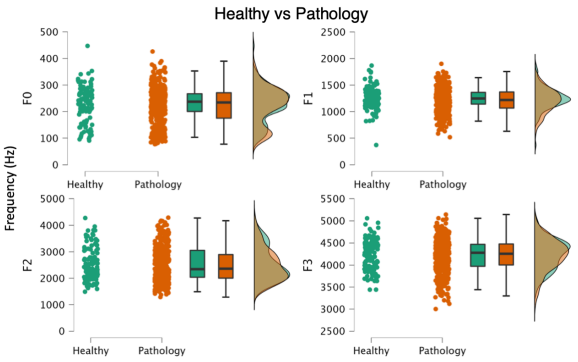


Figure 1: *Distribution of F0, F1, F2 and F3 in healthy and pathology groups.*

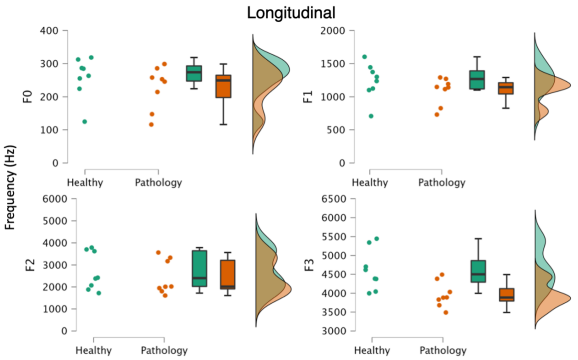


Figure 2: *Distribution of F0, F1, F2 and F3 for the same 8 subjects during pathology and after recovering (healthy).*

F3), preventing any wider generalization. An earlier examination of formants extracted from /ɑ:/ for children of various ages with asthmatic symptoms revealed that asthmatic formants were lower in some age groups than in their healthier counterpart age group [4]. Though that conclusion follows the findings of this study, it is important to highlight that the number of cases examined in that study [4] was quite small (only 6 to 14 different children in each age group) and the comparison was limited to asthma with no hypothesis testing thus cautioning against any gross generalization.

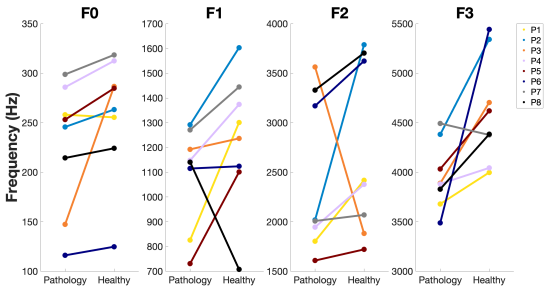


Figure 3: *F0, F1, F2 and F3 feature evolution for the same 8 subjects during pathology and after recovering (healthy).*

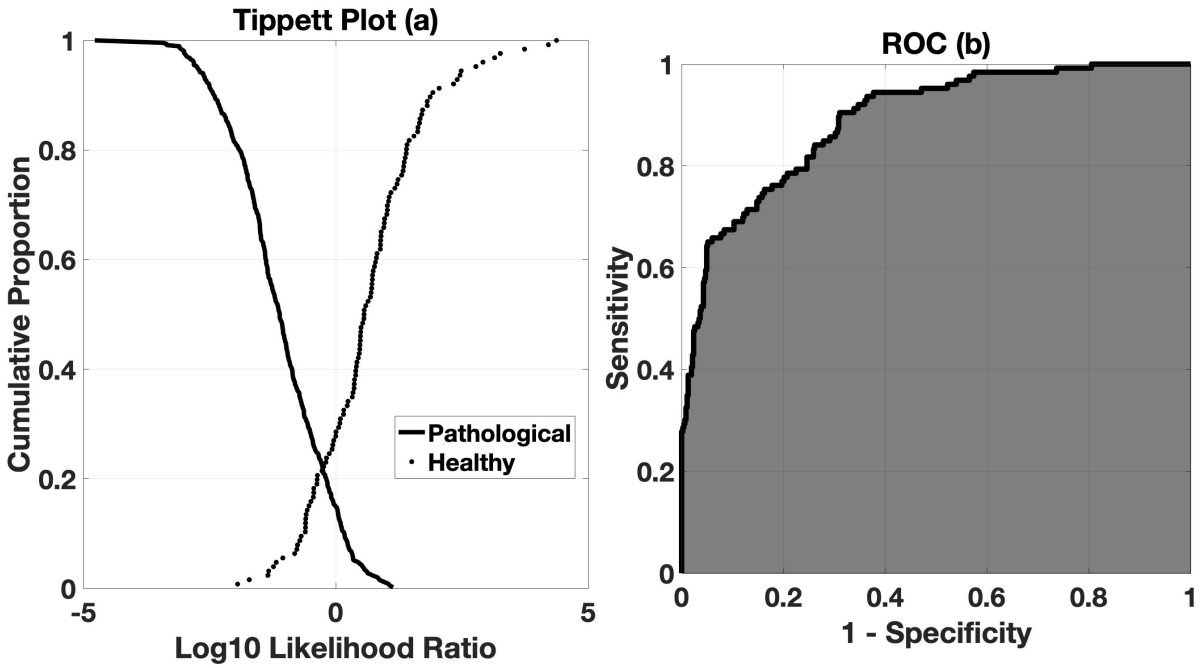


Figure 4: *(a) Tippett plots and (b) ROC of model trained using vocalised /ɑ:/ sounds.*

5.2. Classification Model Results

Table 4 shows the model's performance for /ɑ:/ sounds in terms of classification accuracy, sensitivity, specificity, and AROC. The healthy and pathological classification was based on the optimal threshold that maximizes the sensitivity and specificity values. The classification accuracy of the model trained using the /ɑ:/ sound was high (81.92%), indicating that the /ɑ:/ sound must contain crucial cues in discriminating a healthy respiratory tract from a pathological one. The model's sensitivity (75.40%) and specificity (83.73%) are also high. Despite the fact that there was no significant difference in the mean values of the F0, F1, F2, and F3 distributions retrieved from healthy and pathological populations, the model resulted in high classification accuracy when trained using automatic audio features. This is unsurprising, given that MFCCs focus on perceptually relevant aspects of the audio spectrum and have proven to be particularly effective in a range of audio classification tasks [22, 23].

Table 4: *Model Performance.*

Attribute	Result
Accuracy	81.92
Sensitivity	75.40
Specificity	83.73
AROC	0.89

Figure 4(a) shows the Tippett plot for the resultant cumulative (uncalibrated) LLR values. The relative symmetry of the plot around the $LLR = 0$ line confirms the unbiased results of the /ɑ:/ sound model. The crossover point between the healthy subject curve and the pathological curve was also found to be low (around 0.2) indicating a relatively low rate of misclassification.

The receiver operating characteristic of this model is shown in Figure 4(b). ROC curve occupies the upper left corner and the resulting AROC value is 0.89. The AROC is convincingly high, which means that the model has delivered good separability between healthy and pathological class.

6. Conclusions

We gathered a unique dataset of /ɑ:/ sounds from both healthy children and children with respiratory pathology. Despite the fact that there were no significant differences between the mean values of the fundamental and formant frequency (F0, F1, F2, and F3) distributions obtained from healthy and pathological populations, a GMM-UBM model using MFCCs extracted from /ɑ:/ was nonetheless still able to achieve a classification accuracy exceeding 82%. This accuracy is particularly impressive given the "raw" in-situ settings of data collection, recorded in-clinic on a simple smartphone. The developed model would have potential in supporting clinical diagnostic assessment, enhancing preoperative screening of paediatric respiratory symptoms and informing clinicians' decisions. Therefore, it invites investigation whether training with more data could enhance the accuracy of /ɑ:/ sound models, given the non-invasive and convenient nature of such a mode of patient assesment. We plan to gather more data in future studies in order to employ deep learning techniques, enhancing performance even further.

7. References

[1] M. Todokoro, H. Mochizuki, K. Tokuyama, and A. Morikawa, "Childhood cough variant asthma and its relationship to classic asthma," *Annals of Allergy, Asthma & Immunology*, vol. 90, no. 6, pp. 652–659, 2003.

[2] A. B. Chang, "Cough, cough receptors, and asthma in children," *Pediatric pulmonology*, vol. 28, no. 1, pp. 59–70, 1999.

[3] H. I. Hee, B. Balamurali, A. Karunakaran, D. Herremans, O. H. Teoh, K. P. Lee, S. S. Teng, S. Lui, and J. M. Chen, "Development of machine learning for asthmatic and healthy voluntary cough sounds: A proof of concept study," *Applied Sciences*, vol. 9, no. 14, p. 2833, 2019.

[4] B. BT, H. I. Hee, O. Teoh, K. Lee, S. Kapoor, D. Herremans, and J.-M. Chen, "Asthmatic versus healthy child classification based on cough and vocalised/:/sounds," *The Journal of the Acoustical Society of America*, vol. 148, no. 3, pp. EL253–EL259, 2020.

[5] S. Yadav, M. Keerthana, D. Gope, P. K. Ghosh *et al.*, "Analysis of acoustic features for speech sound based classification of asthmatic and healthy subjects," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6789–6793.

[6] K. D. Bartl-Pokorny, F. B. Pokorny, A. Batliner, S. Amiriparian, A. Semertzidou, F. Eyben, E. Kramer, F. Schmidt, R. Schönweiler, M. Wehler *et al.*, "The voice of covid-19: Acoustic correlates of infection in sustained vowels," *The Journal of the Acoustical Society of America*, vol. 149, no. 6, pp. 4377–4383, 2021.

[7] M. Asiaee, A. Vahedian-Azimi, S. S. Atashi, A. Keramatfar, and M. Nourbakhsh, "Voice quality evaluation in patients with covid-19: An acoustic analysis," *Journal of Voice*, 2020.

[8] Y. Amrulloh, U. Abeyratne, V. Swarnkar, and R. Triasih, "Cough sound analysis for pneumonia and asthma classification in pediatric population," in *2015 6th International Conference on Intelligent Systems, Modelling and Simulation*. IEEE, 2015, pp. 127–131.

[9] S. Yadav, N. Kausthubha, D. Gope, U. M. Krishnaswamy, and P. K. Ghosh, "Comparison of cough, wheeze and sustained phonations for automatic classification between healthy subjects and asthmatic patients," in *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2018, pp. 1400–1403.

[10] K. S. Alqudaihi, N. Aslam, I. U. Khan, A. M. Almuhaideb, S. J. Alsunaidi, N. M. A. R. Ibrahim, F. A. Alhaidari, F. S. Shaikh, Y. M. Alsenbel, D. M. Alalharith *et al.*, "Cough sound detection and diagnosis using artificial intelligence techniques: challenges and opportunities," *Ieee Access*, vol. 9, pp. 102 327–102 344, 2021.

[11] S. A. H. Tabatabaei, P. Fischer, H. Schneider, U. Koehler, V. Gross, and K. Sohrabi, "Methods for adventitious respiratory sound analyzing applications based on smartphones: a survey," *IEEE reviews in biomedical engineering*, vol. 14, pp. 98–115, 2020.

[12] V. Nathan, M. M. Rahman, K. Vatanparvar, E. Nemati, E. Blackstock, and J. Kuang, "Extraction of voice parameters from continuous running speech for pulmonary disease monitoring," in *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, 2019, pp. 859–864.

[13] I. Song, "Diagnosis of pneumonia from sounds collected using low cost cell phones," in *2015 International joint conference on neural networks (IJCNN)*. IEEE, 2015, pp. 1–8.

[14] A. Murata, Y. Taniguchi, Y. Hashimoto, Y. Kaneko, Y. Takasaki, and S. Kudoh, "Discrimination of productive and non-productive cough by sound analysis," *Internal Medicine*, vol. 37, no. 9, pp. 732–735, 1998.

[15] A. A. Abaza, J. B. Day, J. S. Reynolds, A. M. Mahmoud, W. T. Goldsmith, W. G. McKinney, E. L. Petsonk, and D. G. Frazer, "Classification of voluntary cough sound and airflow patterns for detecting abnormal pulmonary function," *Cough*, vol. 5, no. 1, p. 8, 2009.

[16] U. R. Abeyratne, V. Swarnkar, A. Setyati, and R. Triasih, "Cough sound analysis can rapidly diagnose childhood pneumonia," *Annals of biomedical engineering*, vol. 41, no. 11, pp. 2448–2462, 2013.

[17] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted gaussian mixture models," *Digital signal processing*, vol. 10, no. 1-3, pp. 19–41, 2000.

[18] G. S. Morrison, "Forensic voice comparison and the paradigm shift," *Science & Justice*, vol. 49, no. 4, pp. 298–308, 2009.

[19] P. Rose, *Forensic speaker identification*. CRC Press, 2003.

[20] L. R. Rabiner and R. W. Schafer, *Theory and applications of digital speech processing*. Pearson Upper Saddle River, NJ, 2011, vol. 64.

[21] L. R. Rabiner and B.-H. Juang, *Fundamentals of speech recognition*. PTR Prentice Hall Englewood Cliffs, 1993, vol. 14.

[22] B. BT, K. Lin, S. Lui, J. Chen, and D. Herremans, "Towards robust audio spoofing detection: a detailed comparison of traditional and learned features," *IEEE Access*, vol. 7, pp. 84 229–84 241, 2019.

[23] L. Muda, M. Begam, and I. Elamvazuthi, "Voice recognition algorithms using mel frequency cepstral coefficient (mfcc) and dynamic time warping (dtw) techniques," *arXiv preprint arXiv:1003.4083*, 2010.

[24] B. B. Nair, E. A. Alzqhoul, and B. J. Guillemin, "Comparison between mel-frequency and complex cepstral coefficients for forensic voice comparison using a likelihood ratio framework," in *Proceedings of the World Congress on Engineering and Computer Science, San Francisco, USA*, 2014.

[25] D. Meuwly and A. Drygajlo, "Forensic speaker recognition based on a bayesian framework and gaussian mixture modelling (gmm)," in *2001: A Speaker Odyssey-The Speaker Recognition Workshop*, 2001.

[26] C. D. Brown and H. T. Davis, "Receiver operating characteristics curves and related decision measures: A tutorial," *Chemometrics and Intelligent Laboratory Systems*, vol. 80, no. 1, pp. 24–38, 2006.

[27] "A self-study guide for aspiring machine learning practitioners," https://developers.google.com/machine-learning/crash-course/, accessed: 2022-04-15.

[28] B. L. Welch, "The generalization of 'student's'problem when several different population varlances are involved," *Biometrika*, vol. 34, no. 1-2, pp. 28–35, 1947.

[29] P. Boersma, "Praat: doing phonetics by computer [computer program]," *http://www. praat. org/*, 2011.

[30] A. W. Toga, *Brain mapping: An encyclopedic reference*. Academic Press, 2015.

Author Index