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

4–7 December 2018 ▪ Sydney, Australia

Conference Information
Table of Contents
Author Index
Search

Editors: Julien Epps, Joe Wolfe, John Smith and Caroline Jones

ASSTA
The Australasian Speech Science & Technology Association Incorporated

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

Glorious Coogee, image copyright Heidi Hereth

Welcome to Delegates

On behalf of the organising committee and the Australasian Speech Science and Technology Association (ASSTA), we have great pleasure in welcoming you to Coogee Beach for the *17th Australasian International Speech Science and Technology Conference – SST 2018.* SST is the premier Australasian and international forum for research into all areas of speech science and technology. The conference focuses on theoretical and empirical foundations, and technologies and applications that define the field.

For SST 2018, 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 expert reviewers, the submissions were all blind peer reviewed by at least two anonymous reviewers (in most cases three), and papers were selected on the basis of reviewer comments and scores. Authors resubmitted final deanonymised versions of their papers and abstracts that took into account the reviewers' comments.

The program was formed by grouping papers into main topics of interest for this year's conference. The conference proceedings have been distributed for download by all participants on the conference website.

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 delegates in 2018 are from overseas, hailing from Brazil, Finland, Germany, Japan, the Netherlands, Poland, Saudi Arabia, Singapore and the USA. We are delighted to welcome our SST 2018 international keynote presenters Kristiina Jokinen and Thomas Quatieri and international tutorial presenters Emily Mower Provost, Martijn Wieling and Kong Aik Lee, as well as local keynote presenters Katherine Demuth and local tutorial presenter Helen Fraser, and we thank them for contributing to the continuing success of our conference series. We are also pleased to welcome special guest Mal Webb, to provide a live demonstration of the capabilities and properties of the vocal apparatus.

We would like to acknowledge UNSW Sydney for its support of the conference, in particular the Schools of Physics and Electrical Engineering and Telecommunications, who provided both financial and logistical support. We also gratefully acknowledge the financial support of the NSW Government.

The chairs 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. Due to the single-track schedule of the conference, all delegates will have ample opportunities to take advantage of the multidisciplinary

nature of the conference, and to hold discussions with speech researchers from other discipline areas.

We hope that you will find this program interesting, engaging, and a stimulus to your next research advances.

Julien Epps
SST'18 Chair
UNSW Sydney

Joe Wolfe
SST'18 Chair
UNSW Sydney

John Smith
SST'18 Chair
UNSW Sydney

Caroline Jones
SST'18 Chair
Western Sydney Univ.

SST 2018 Sponsors

Host organisation
Sponsored and organised by
School of Physics
School of Electrical Engineering and Telecommunications

Generously sponsored by the NSW Government 2108
Conference Sponsorship Program

SST 2018 Conference Organization

General Chairs: Julien Epps *(UNSW Sydney)*
Joe Wolfe *(UNSW Sydney)*
John Smith *(UNSW Sydney)*
Caroline Jones *(Western Sydney University)*

Web Chair: Zhaocheng Huang *(UNSW Sydney)*

Reviewers:

Waleed Abdulla
Beena Ahmed
Brett Baker
Titia Benders
Tessa Bent
Rosey Billington
Andre Goios Borges de Almeida
Laurence Bruggeman
Rikke Bundgaard-Nielsen
Sasha Calhoun
Yan Chen
Frantz Clermont
Felicity Cox
Ting Dang
Katherine Demuth
Gerry Docherty
Ewald Enzinger
Paola Escudero
Christopher Fennell
Janet Fletcher
Paul Foulkes
David Grayden
Bernard Guillemin
John Hajek
Noel Hanna
Jonathan Harrington
Mark Harvey
Yusuke Hioka
Colleen Holt
Zhaocheng Huang
Vincent Hughes
Saad Irtza
Shunichi Ishihara
Nenagh Kemp
Yuko Kinoshita
Carmen Kung
Trent Lewis
Debbie Loakes
Olga Maxwell
Geoffrey Morrison
Mitsuhiro Nakamura
Sven Nordholm
Michael Proctor
Michael Robb
Phil Rose
Iris-Corinna Schwarz
Belinda Schwerin
Vidhyasaharan Sethu
Stephen So
Kaavya Sriskandaraja
Hywel Stoakes
Marija Tabain
Roberto Togneri
Kimiko Tsukada
Chiharu Tsurutani
Michael Tyler
Adam Vogel
Michael Wagner
Petra Wagner
Catherine Watson
Stephen Winters
Maria Wolters
Janice Wing Sze Wong
Nan Xu
Ivan Yuen
Cuiling Zhang

Seventeenth Australasian International Conference on Speech Science and Technology

SST 2018

Table of Contents

Analysis, Processing and Forensics 2

Infant-Directed Speech

Acoustic Phonetics

Regional Languages and Variants 2

Regional Languages and Variants 3

Analysis, Processing and Forensics 3

Phonetics and Linguistics

Regional Languages and Variants 4

Analysis, Processing and Forensics 4

Poster Session

A study of formants preceding apical consonants in Pitjantjatjara

Marija Tabain[1], Richard Beare[2, 3]

[1] Department of Languages and Linguistics, La Trobe University, Melbourne Australia
[2] Department of Medicine, Monash University, Melbourne, Australia
[3] Murdoch Children's Research Institute, Melbourne, Australia

m.tabain@latrobe.edu.au, richard.beare@ieee.org

Abstract

This study presents formant data for apical consonants from three female speakers of Pitjantjatjara. Formants are extracted from the vowel preceding the six apical consonants of the language: alveolar /t n l/ and retroflex /ʈ ɳ ɭ/. Results show that there are almost no formant differences between alveolars and retroflexes when the preceding vowel is /i/. By contrast, effects are found for the other two vowels of the language /a/ and /u/. The difference between F3 and F2 is smaller for retroflexes, and larger for post-tonic alveolar consonants. Limited effects are also found on F1 and F4.

Index Terms: alveolar, retroflex, apical, stop, nasal, lateral, Australian languages.

1. Introduction

Pitjantjatjara is a dialect of the greater Western Desert language of Central Australia, and is spoken several hundred kilometres south-west of the administrative township of Alice Springs, towards Ayers Rock/Uluṟu and the border between the Northern Territory and South Australia [1, 2, 3]. It has around 2000 speakers and is still being learned by children today.

Like many Australian languages, Pitjantjatjara has a relatively rich series of coronal consonant contrasts [4, 5]. Pitjantjatjara has alveolar, retroflex and alveo-palatal places of articulation, at each of the stop, nasal and lateral manners of articulation. The focus of the present study is on the alveolar versus retroflex contrast in Pitjantjatjara: /t ʈ/, /n ɳ/ and /l ɭ/. Both the alveolar and retroflex places of articulation are classed as apical articulations (different from the laminal articulation of the alveo-palatal consonants).

Pitjantjatjara maintains the alveolar versus retroflex contrast for all three phonemic vowels of the language /a i u/ (there is also a phonemic vowel length contrast, which we ignore here). The apical contrast is maintained in hetero-organic consonant clusters (e.g. /wanka/ *awake, alive* versus /waŋka/ *spider* – c.f. /waŋka/ *talk!*) and in the typologically infrequent /i/-vowel context (e.g. /ini/ *name* versus /iɳi/ *loose, shaky*).

However, as is typical of the alveolar versus retroflex contrast in the world's languages, [6, 7, 8] – but see [9], this apical contrast is neutralized in word-initial position in Pitjantjajtara. In the present study, we do not consider the neutralized apical (which can be represented as /T/, /N/ or /L/ for the different manners of articulation respectively), and instead focus on the apicals that are contrastive, in all three vowel environments.

Previous acoustic work on Pitjantjatjara apical stops, focusing on the stop burst at the right edge of the consonant [10, 11], has shown very little difference between the (non-neutralized) alveolar and retroflex. However, the neutralized initial apical was shown to have a higher spectral centre of gravity than either of the non-neutralized apicals. Moreover, various measures of the non-neutralized stop burst suggested that the release of the retroflex /ʈ/ preceding an /i/ vowel was at a more anterior location than the release of the alveolar /t/. This result was interpreted as being due to articulatory overshoot of the target for /ʈ/ before /i/, perhaps due to the difficulty of combining tongue-tip retraction with a high palatal gesture for the following vowel. This result provided additional explanation for the cross-linguistic rarity of the retroflex in the /i/ vowel context.

In the present study we focus on the formants in the vowel preceding the apical consonant, since the preceding vowel transition is deemed to provide the main cue to the alveolar versus retroflex contrast [7]. In addition, we examine the contrast in light of the prosodic factor of lexical stress. This focus is motivated by articulatory data from the neighbouring language Arrernte, where it has been shown that although the apical contrast is quite marginal, with much overlap between the two phonemic apical categories [12], the contrast is in fact mediated by lexical stress. The most prototypical retroflex articulation – involving closure in the post-alveolar or pre-palatal region, followed by a ballistic forward movement of the tongue during consonant closure – was found to occur following the stressed vowel of Arrernte. By contrast, the most prototypical alveolar articulation – with both closure and release of the consonant occurring at the alveolar place of articulation – was more likely to occur preceding the stressed vowel. However, many intermediate articulations were also observed, leading to much variability overall across the two categories.

In the present study of Pitjantjajtajra, we thus examine the apical contrast according to whether the previous vowel is stressed (here termed "post-tonic" position) or whether the previous vowel is unstressed (here termed "weak" position). It has previously been noted that post-tonic position is a "strong" prosodic position in Australian Aboriginal languages [14], with consonant contrasts being maximized in this environment. The apical preceding the stressed vowel is not studied here, since this is the initial neutralized apical /T/ ~ /N/ ~ /L/ (in Pitjantjatajra stress occurs on the first syllable/vowel of the word [13]).

In the present study, we also extend the previous work on Pitjantjatjara by looking not only at apical stops, but also at apical sonorants – i.e. nasals and laterals – in order to gauge how robust the apical contrast is for different manners of articulation.

2. Method

2.1. Speakers and Recordings

Data are presented for 3 female speakers of Pitjantjatjara from Areyonga community in the Northern Territory. Data were recorded direct to computer in a professional-grade recording studio at LaTrobe University in 2010, under the supervision of a professional recording technician. The recordings were acquired in WAV format at a sample rate of 44.1 kHz with 32 bits per sample.

Stimuli consisted of single words which were repeated by the speaker three times in a row, without carrier phrase. The list of words was designed to illustrate the sounds of the language in different positions in the word (i.e. word-initial, word-medial, and, where possible, word-final position), and in different vowel contexts. Dis-fluent tokens were discarded.

2.2. Analyses

Apical stop, nasal and lateral tokens were extracted for analysis. Tokens were either intervocalic, or word-final. Although all three vowel contexts /a i u/ were included, the vowel contexts surrounding the consonant were not necessarily symmetrical (e.g. although /ata/ is a possible sequence extracted for the present study, so are /ati/ and /atu/). Tokens were coded according to whether the preceding vowel was lexically stressed or not: if the preceding vowel was stressed, the token was labelled "Post-Tonic"; and if the preceding vowel was unstressed, the token was labelled "Weak". (Word-initial tokens are not included in the present study: these would be classified as pre-tonic, i.e. before the stressed vowel in the word, and as noted above would consist only of the neutralized /T/ ~ /N/ ~ /L/).

All analyses were conducted using the EmuR speech analysis software [15, 16] interfaced with the R statistical package version 3.4.3 [17, 18]. Vowel formants were extracted based on the Snack pitch and formant tool in the legacy Emu speech software (with default settings except for frame-rate which was set to 200 Hz). Formants were sampled 10 ms before the end of the preceding vowel. This strategy was chosen because there are sudden changes in cavity affiliation of resonances at the boundary between a vowel and a nasal consonant, or between a vowel and a lateral consonant (c.f. [19]). By sampling 10 ms before the vowel-consonant boundary, it is expected that only vowel formants will be sampled, rather than consonant formants. It is of course the case that by sampling at a point that is not perfectly located at the acoustic start or endpoint of the vowel, the most extreme value of any formant movement will not be captured – this is a recognized limitation of the strategy designed to minimize formant-tracking errors. In addition, any tokens which contained formant values which could be considered to be clear tracking errors were removed from analysis, and only tokens which contained formant values within the following ranges were included: F1 200-800 Hz; F2 800-2800 Hz; F3 1500-4000 Hz; and F4 2500-4800. These ranges were chosen based on visual inspection of the data.

2.3. Number of Tokens

Table 1 shows the number of stop, nasal and lateral tokens extracted from the database (after removal of formant tracking errors). The tokens are divided according to whether they are alveolar or retroflex; post-tonic or weak. There are thus four categories: alveolar post-tonic, alveolar weak, retroflex post-tonic, and retroflex weak.

Table 1. *Number of tokens.*

	alveolar post-tonic	alveolar weak	retroflex post-tonic	retroflex weak	
stop	196	258	306	150	**910**
nasal	215	347	173	267	**1002**
lateral	305	159	321	150	**935**
	716	**764**	**800**	**567**	**2847**

It can be seen that there are relatively even numbers of tokens in each manner of articulation, although there are slightly more nasal tokens (about 1000 nasal tokens, compared to about 900 stop and 900 lateral tokens). The least frequent prosodic category in this database is the retroflex weak category (about 500 tokens, as opposed to about 700-800 tokens in the other prosodic categories). However, subtle differences exist according to manner: for the stops, alveolar weak and retroflex post-tonic are more common than alveolar post-tonic and retroflex weak; for the nasals, weak tokens are more common than post-tonic tokens; and for the laterals, post-tonic tokens are more common than weak tokens.

Table 2. *Mean consonant duration (in milliseconds).*

	alveolar post-Tonic	**alveolar weak**	**retroflex post-Tonic**	**retroflex weak**
stop	106	90	86	80
nasal	146	85	135	79
lateral	105	53	94	57

Table 2 shows duration of consonant closure for the apical consonants in this study. It can be seen that the different manners of articulation show slightly different durational patterns. For the nasals, the post-tonic consonants have a duration of around 140 ms, while the prosodically weak consonants have a duration of around 80 ms. For the laterals, by contrast, the post-tonic consonants have a duration of around 100 ms, while the prosodically weak consonants have a duration of around 55 ms (i.e. the laterals are overall shorter in duration, but maintain a similar prosodic contrast, with weak tokens being about half the duration of post-tonic consonants). The stop consonants have a less robust contrast according to prosodic category, with all durations tending around 80-110 ms: although prosodically weak stops have a shorter duration than their post-tonic counterpart, the difference in duration is not as great as for the sonorant consonants. Nevertheless, Table 2 confirms that the contrast between post-tonic and weak apical consonants is reliably encoded in terms of duration.

3. Results

Figure 1 shows the formants preceding the apical consonants. For each formant, nine panels are shown: the rows show data according to manner of articulation (stop, nasal, lateral) and the columns show data according to the preceding vowel /a i u/. Note that there are no weak retroflex nasal or lateral tokens which have a preceding /i/.

If can be seen that overall, there are relatively few differences in F1 for the apical consonants. The clearest difference can be found for the stop consonants in the /a/

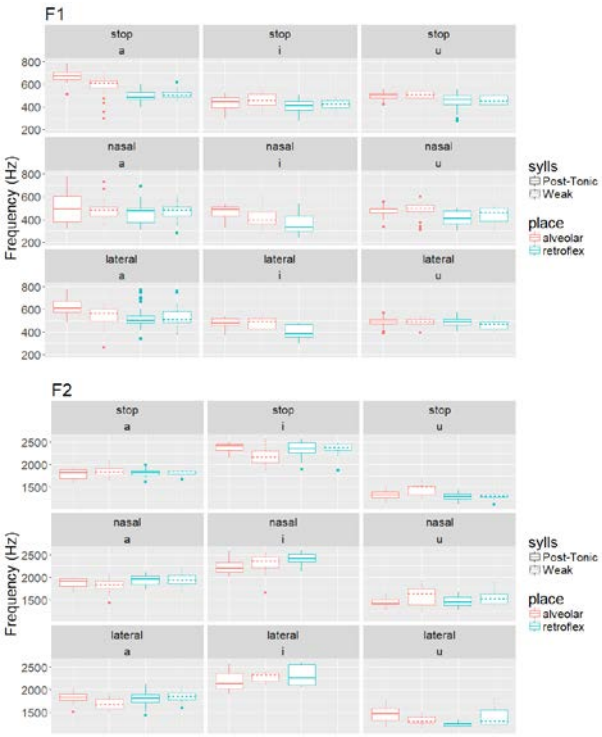
context, where F1 is lower preceding a retroflex consonant, and highest preceding a post-tonic alveolar (top left panel). There are hints of the same pattern for the other manners and for the other vowel contexts – in particular the lower F1 preceding a retroflex consonant – although examination of individual speaker data showed that the pattern was not consistent across all three speakers. The least well-differentiated apical contrast in F1 was for the stops following an /i/ vowel (centre top panel).

Results for F2 are similar, in that there are relatively few differences between the apical consonants. Although there is some evidence for a slightly lower F2 preceding the alveolar weak consonants, examination of individual speaker plots shows much inter-speaker variation overall.

F3, by contrast, shows some clearer differences between the apicals. F3 is lower in the /a/ context preceding the retroflexes for all three manners of articulation. In the case of laterals, F3 is lowest for the post-tonic retroflex. F3 is also noticeably lower in the context of /u/, particularly for the stop and nasal manners. For the lateral, the weak alveolar patterns with the retroflexes (although there is also some evidence that F3 is higher preceding the post-tonic alveolars, e.g. the nasal in an /a/ context, as well as the lateral in an /u/ context).

In the context of /i/, there are almost no differences between the various apicals, although the weak retroflex has a slightly lower F3 for the stop.

An examination of F4 data shows some differences between the manners of articulation. The stop consonants show a lower F4 preceding the retroflex stops in the context of /a/ and /u/, but not in the context of /i/. These particular results echo the F3 results. However, the differences are not so clear for the nasal and lateral manners of articulation, although there are occasional hints of a higher F4 preceding post-tonic alveolars (e.g. in the /i/ context), or slightly lower F4 for the retroflexes in the other vowel contexts /a/ and /u/.



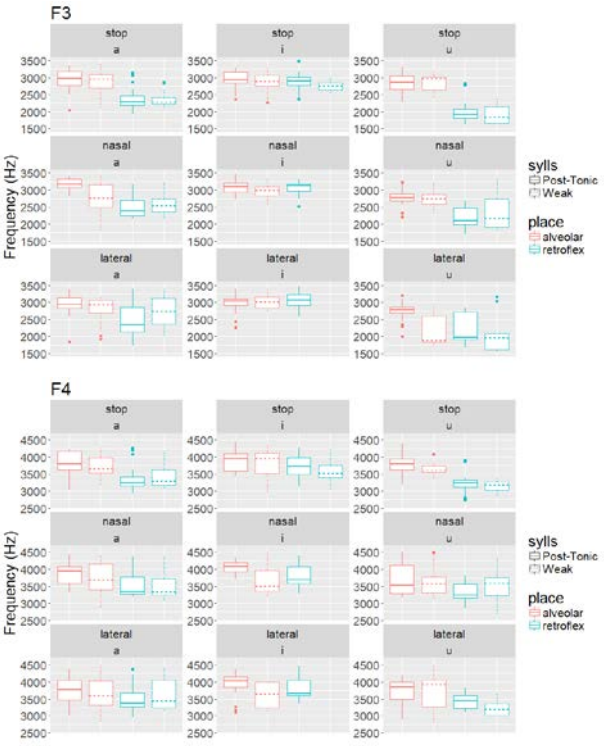


Figure 1. *Formants 1-4 averaged across speakers.*

Finally, Figure 2 shows the difference between F3-F2. A much clearer pattern of results emerges when data are presented to take this spectral prominence into account. A clear separation between alveolars and retroflexes is evident for all three manners of articulation in the context of /a/ (left column). This is also true in the context of /u/, with the exception of the lateral, where the weak alveolar patterns with the retroflexes (bottom right panel). There is some evidence of a greater difference in F3-F2 for the post-tonic alveolar, in particular the nasal in the /a/ context, and the nasal and the lateral in the /u/ context); and perhaps also a smaller difference in F3-F2 for the post-tonic retroflex (e.g. the nasal and the lateral in the /a/ context). However, these effects are not strong, and are perhaps best considered a trend pending analysis from further speakers.

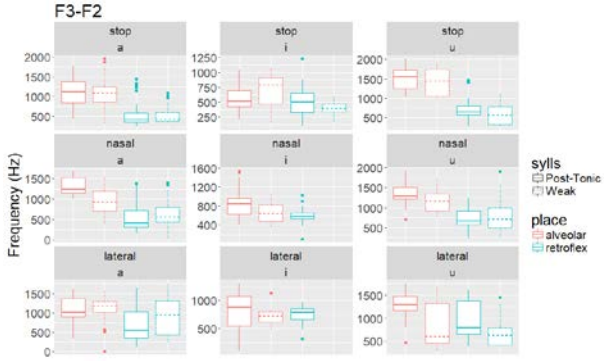


Figure 2. *F3-F2 averaged across speakers.*

4. Discussion

The results presented above confirm that the difference between F3 and F2 is the clearest cue to apical place of articulation. Although the data above suggest that much of this difference is due to a lower F3 in the case of retroflexes, a closer examination of individual speaker results shows that for some speakers, this small F3-F2 difference is also due to a slightly higher F2. The above results also suggest that a lower F4 may play a role in retroflex identity, at least for the stop consonants; and a lower F1 may also play a role in retroflex identity, at least for the stop consonants in the context of /a/.

The above formant results confirm previous observations that retroflexion is difficult in the context of an /i/ vowel. The alveolars and retroflexes were not differentiated at all when the preceding vowel was /i/, at least not in terms of the formants examined here. This suggests that retroflexion is not compatible with tongue-body fronting, and the previous results from spectral bursts suggest that this incompatibility leads to eventual overshoot of the retroflex release target. By contrast, there were much stronger formant differences between alveolars and retroflexes in the context of preceding /a/ and /u/, which suggests that tongue backing does not interfere with retraction of the tongue tip for retroflexion.

Manner differences became more evident when examining the higher formants. Contrast between the alveolar and retroflex was much clearer for the stops than for the nasals and laterals when F3 and F4 were examined. This is despite the fact that vowel formants were sampled 10 ms from the marked boundary between consonant and vowels, and thus where damping effects from the following sonorant would be expected to be minimal. This result is also despite acoustic and articulatory evidence that Australian languages align velo-pharyngeal closing/opening very tightly with supra-laryngeal closing/opening, precisely in order to minimize the vowel nasalization which would compromise acoustic cues to place of articulation [20]. The above results for manner contrasts clearly require further consideration, although they do suggest that stops allow for the clearest acoustic cues to consonant place of articulation.

As regards the prosodic contrasts examined in the present study, there was little evidence of a "stronger" retroflex in post-tonic position (at least as measured here in the formants), as might have been expected based on the articulatory data from Arrernte cited above. However, for the three speakers in the current study at least, there was evidence for a more prototypical alveolar articulation in post-tonic position, for instance with a higher F1 in the /a/ context for the stop (suggesting a lower vowel position preceding the alveolar), and a higher F3-F2 value (i.e. a greater difference between these two formants) in the /a/ and /u/ context for the nasal and /u/ for the lateral. This suggests that the apical contrast is indeed more likely to be clearly realized in post-tonic position, although there is still a good deal of overlap between the two phonemic apical categories. This last observation provides some support for the notion that the apical contrast is marginally phonemic even in Pitjantjatjara, since in the weak prosodic context, the alveolar does occasionally pattern with the retroflexes in terms of preceding formant structure.

5. Acknowledgements

We would like to thank the Australian Research Council for funding this research. We would also like to thank our speakers – Hilda Bert, Charmaine Coulthard and Kathleen Windy – for their commitment to speech research. *Palya*

6. References

[1] C. Goddard, *Pitjantjatjara/Yankunytjatjara to English Dictionary* (2nd edition). Alice Springs: IAD Press, 1996.

[2] C. Goddard, *A Learner's Guide to Pitjantjatjara/Yankunytjatjara.* Alice Springs: IAD Press, 1993.

[3] M. Tabain & A. Butcher, "Pitjantjatjara: Illustration of the IPA", *Journal of the International Phonetic Association* **44** 189-200, 2014.

[4] R. Dixon, *Australian Languages: Their Nature and Development.* Cambridge: Cambridge University Press, 2002.

[5] N. Evans, "Current issues in the phonology of Australian languages," In John Goldsmith (Ed.), *The Handbook of Phonological Theory* pp. 723–761). Cambridge, MA, Oxford: Blackwell, 1995.

[6] H. Simonsen, I. Moen & S. Cowen, "Norwegian retroflex stops in a cross-linguistic perspective," *Journal of Phonetics*, vol. 36, pages 385–405, 2008.

[7] D. Steriade, "Directional asymmetries in place assimilation: A perceptual account," In E. Hume & K. Johnson (Eds.) *The role of speech perception in phonology* (pp. 219–250). San Diego: Academic Press, 2001.

[8] A. Kochetov and N. Sreedevi, "Manner-specific tongue shape differences in the production of Kannada coronal consonants," paper presented at the Spring 2016 meeting of the Acoustical Society of America, 2016.

[9] R. Bundgaard-Nielsen, B. Baker, C. Kroos, M. Harvey & C. Best. "Vowel acoustics reliably differentiate three coronal stops of Wubuy across prosodic contexts", *Laboratory Phonology* **29** 133–161, 2012.

[10] M. Tabain & A. Butcher, "Stop bursts in Pitjantjatjara", *Journal of the International Phonetic Association* **45** 149-176, 2015.

[11] M. Tabain & A. Butcher, "Lexical stress and stop bursts in Pitjantjatjara: feature enhancement of neutralized apicals and the coronal/velar contrast", *Journal of Phonetics* **50** 67-80, 2015.

[12] M. Tabain, "An EPG study of the alveolar vs. retroflex apical contrast in Central Arrernte," *Journal of Phonetics,* vol. **37,** pages 486-501, 2009.

[13] M. Tabain, J. Fletcher & A. Butcher, "Lexical stress in Pitjantjatjara", *Journal of Phonetics* **42** 52-66, 2014.

[14] A. Butcher. Australian Aboriginal Languages: Consonant-Salient Phonologies and the 'Place-of-Articulation Imperative'. In J. Harrington & M. Tabain (eds). *Speech Production: Models, Phonetic Processes, and Techniques.* New York, USA: Psychology Press, pp. 187-210, 2006.

[15] R. Winkelmann, J. Harrington & K. Jänsch.. EMU-SDMS: Advanced speech database management and analysis in R. *Computer Speech and Language* **45** 392-410, 2017.

[16] J. Harrington, *The Phonetic Analysis of Speech Corpora.* Blackwell, 2010.

[17] R Core Team "R: A language and environment for statistical computing. R Foundation for Statistical Computing," Vienna, Austria. URL http://www.R-project.org/, 2014.

[18] H. Wickham. *ggplot2: Elegant Graphics for Data Analysis.* Springer-Verlag New York, 2009.

[19] K. Stevens. *Acoustic phonetics.* Cambridge MA: MIT Press, 1998.

[20] H. Stoakes. An acoustic and aerodynamic analysis of consonant articulation in Bininj Gun-wok. PhD thesis, University of Melbourne, 2014.

Sociophonetic variability of post-vocalic /t/ in Aboriginal and mainstream Australian English

Debbie Loakes[1, 3], Kirsty McDougall[2], Joshua Clothier [1,3], John Hajek[1] and Janet Fletcher[1,3]

[1] The University of Melbourne
[2] The University of Cambridge
[3] ARC Centre of Excellence for the Dynamics of Language

dloakes@unimelb.edu.au, kem37@cam.ac.uk, {joshuajc, j.hajek, j.fletcher}@unimelb.edu.au

Abstract

This paper analyses post-vocalic /t/ variability in controlled speech across two groups, both L1 Aboriginal English and mainstream Australian English speakers. Data were collected in Warrnambool, a small community in western Victoria (Australia). While both Aboriginal English and mainstream Australian English speakers used canonical aspirated [t ʰ] a range of other variants were observed. The Aboriginal English group used a greater number of variants overall, and tended toward "glottal" variants (full glottal stops, pre-glottalised stops, and ejective-like stops) whereas the mainstream Australian group preferred so-called "breathy" variants (affricates, fricatives); we attribute this to sociophonetic variability, potentially linked with voice quality and glottal timing. Overall, the study highlights some previously undocumented variation both within L1 Aboriginal English, and between L1 Aboriginal English and mainstream Australian English.

Index Terms: plosives, variation, Aboriginal English, mainstream Australian English, sociophonetics

1. Introduction

1.1. Background – /t/ variability in mainstream Australian English

While earlier work on the sociophonetics of Australian English mostly investigated vowel variation, in recent years variation in consonant production has increasingly become a focus of research. For voiceless plosives, and /t/ in particular, phonetic context, speaking style, regional location, sex, accent variety and socioeconomic status have all been noted as playing a part in the patterns of variation exhibited [1,2,3,4].

1.2. L1 Aboriginal English

Aboriginal English is an under-studied variety spoken in Australia. The variety developed in Aboriginal communities, separate from the standard/mainstream [5], and the name *Aboriginal English* is often used as an umbrella term to apply to multiple varieties depending on the location and language experience(s) of the speakers [5,6]. In this study, we use the singular *Aboriginal English* because we are discussing one specific community of first language (L1) speakers (see 2.1). Compared with mainstream Australian English, the "standard" variety in Australia, Aboriginal English is characterised by a somewhat different sound system, as well as differences in every other aspect of language such as grammar, semantics, lexicon and pragmatics [6]. Aboriginal English ranges on a continuum from sounding very similar to standard Australian English ('light' Aboriginal English) to diverging markedly ('heavy'), and speakers themselves can vary depending on the situation and audience [5,6]. Speakers can also be first (L1) or second language (L2) speakers.

1.2.1. Aboriginal English stop consonant production

There is not a great deal of phonetic research into characteristics of Aboriginal English, but there are some detailed summaries of its features. For example [5] provides an overview of Australian Aboriginal English phonology, and while much of what the author describes for stop consonants appears restricted to L2 varieties (i.e. lack of voicing distinctions in non-initial positions, retroflex variants and dental realizations of /t/ in some communities), the statement about the "phonemic boundaries [being] more porous" [5:134] certainly appears to apply, at least impressionistically, to the L1 Aboriginal English analysed in the current study. An overview of Aboriginal English stops is also given in [6] noting similarly that there is variability, but adding more phonetic detail to this picture. In [6] there is a discussion about the fact that Aboriginal English stops may indeed have a voicing contrast initially, but that the contrast may be deployed differently from mainstream Australian English – with the L2 Aboriginal English initial stops tending to be unaspirated. For L2 varieties, wide variability in general, and hypercorrection, are also said to be evident in stop production, and intra-speaker variability is also said to be large [6].

The only study we aware of which specifically addresses phonetic stop consonant production in Aboriginal Englishes is a recent paper [7] focusing on /p t k b d g/ variability between L1 Aboriginal English spoken in Northern Australia and mainstream Australian English spoken in Sydney. In particular, the authors analysed the stop voicing contrast, using Voice Onset Time (VOT), VTT (Voice Termination Time) and duration. They showed that the stop voicing contrast exists for these Aboriginal English speakers, and that there are no significant differences for VOT across the varieties. However, they also found VTT variability in the "voiceless" stop category for the L1 Aboriginal English speakers, whereby many voiceless tokens had some periods of "passive phonetic voicing" [7: 15-16]; this was not observed at all for the mainstream Australian English group. Confirming the prediction in [6], the authors conclude that while both groups use a voicing distinction, they use different strategies. In [7] the extent of variability across L1 and L2 Aboriginal English stops is quantified, and is shown to be statistically much greater than that seen in the mainstream variety.

1.3. Aims

Given that Aboriginal English has developed alongside the mainstream variety but separate from it [5], and that stop production is known to be a domain of extensive (socio)phonetic variability [8], comparing (post-vocalic) /t/ production across Aboriginal and mainstream Australian English from the same small community should highlight any local variability in production behaviour. The aim of the present study is therefore to determine the extent of sociophonetic variability in Aboriginal English and mainstream Australian English spoken in Warrnambool. We predict there will be differences in how speakers across the two groups produce post-vocalic /t/, and we also expect some gendered patterning within speakers from the same groups, given the variation observed in other studies. Aside from being a comparative study, this research will provide a first phonetic description of the extent of fine-grained phonetic variability that occurs in an L1 Aboriginal English variety.

2. Method and Analysis

2.1. Participants and experimental task

This study compares two groups of adult L1 English speakers from Warrnambool, a regional coastal city located in the south west of Victoria. It has a population of approximately 35,000 people, and is located 257 km from Melbourne, approximately three hours away by road.

The participants are 10 Aboriginal English speakers (5 M, 5 F) and 15 mainstream Australian English speakers (8 M, 7 F), recorded in 2015 and 2012 respectively by the first author. The participants were all adults, with wide variability in age. The Aboriginal English group were aged between 19 and 65 years, and the mainstream group was aged between 26 and 72 years. The participants took part in a number of activities (questionnaire, perception test, interview, reading a word list). The data used in the current study are post-vocalic /t/ extracted primarily from word-final contexts in /hVt/ words and some intervocalic tokens from /hVtV/. The sample contains 864 tokens overall (466 for Aboriginal English speakers, 398 for the mainstream Australian English group). The majority of tokens were word final /t/, but 12% of the mainstream Australian English sample consisted intervocalic tokens along with 14% of the mainstream Australian English sample. We consider this a preliminary analysis in the sense that we also have spontaneous speech we wish to analyse, and the Aboriginal English sample is a subset from a larger group.

2.2. Analysis

2.2.1. Phonetic analysis

Speech data were manually-labelled by the first author using *Praat* [10] after autosegmentation of the phonemes using MAUS. The overall quality of each /t/ (and release where present) was visually categorised and identified from spectrograms and annotated on a "phonetic" tier. A tier "t-category" included classification decisions (*canonical, fricative, glottal*, etc. – see below for a full description). The release phase was labelled H in most cases, but also occasionally S which better suited any spirantised or affricate variants. As mentioned, some of this analysis has been presented in [9]. For the mainstream Australian English speaking community, the categories *canonical, affricate, fricative, intermediate* and *tap* accounted for all of the realisations of post-vocalic /t/ in the data set. Some additional categories were necessary when labelling the Aboriginal English data set, namely *glottal, ejective, voiced, approximant*. The following list gives the category names and explanations about the decisions made during the labelling process. This also gives a sense of the acoustic structure of /t/ in the data.

Canonical [tʰ]: period of full closure followed by burst. No voicing apparent [8,11].
Affricate [tˢ]: a closure followed by /s/-like release (not aspirated), no burst like characteristics [e.g. 3].
Fricative [t̪]: a fully fricated variant, not the same as [s], better described as a "lowered /t/" [3].
Intermediate: this category is best described as **[t̪ʰ]**. it has the auditory percept of a fricated stop, but there are burst characteristics evident acoustically [3].
Tap [ɾ]: durationally very short, only observed inter-vocalically [e.g.12].
Approximant [ɾ̞]: technically a tap which did not have full closure, also observed intervocalically [e.g. 12].
Glottal-unreleased (same as **pre-glottalised**) **[ʔt̚]:** these stops have glottal activity and unreleased supralaryngeal closure [4].
Glottal [ʔ]: these are full glottal stops; for these sounds there was no apparent supralaryngeal closure characteristics. These stops can be either plosive-like or creaky in appearance [11], we observed both in our data. [13] note that glottal stops are a possible allophone in mainstream Australian English (phrase-finally), but they must be rare, as they are not observed in our mainstream Australian English sample, nor mentioned in other recent studies focusing on Australian English /t/ [1,2,3,4,9].
Ejective [t']: 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 [13]. In our data there were often sharp "spikes" on the waveform, similar to examples in [14] where this is correlated with burst intensity. In English ejectives may also be considered "emphatic" stops [14].
Voiced [d̬]: these tokens were partially voiced, similar to what [7] describe for their Aboriginal English groups, with passive phonetic voicing in what is a phonologically voiceless category. To our knowledge these have not been observed in mainstream Australian English [e.g. no mention in 4, 12].

2.2.2. Statistical analysis

Statistical analysis was carried out by the third author using *SPSS Statistics 24* on the counts across the Aboriginal and mainstream Australian English samples, and within the Aboriginal sample (analysis within the mainstream sample was carried out in [9]).

3. Results

3.1. Group comparison: distribution of Aboriginal and mainstream Australian English post-vocalic /t/ variants

Figure 1 shows the per-group percentages of the phonetic variants observed in the data comparing the two varieties.

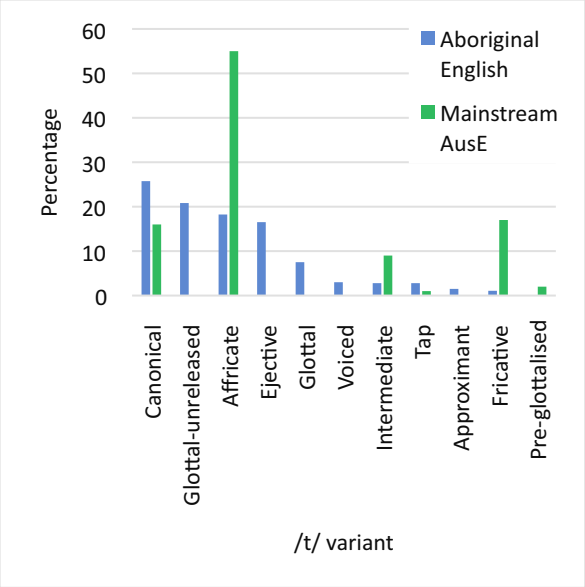


Figure 1. *Distribution (%) of post-vocalic /t/ variants across the Aboriginal and mainstream Australian English samples*

This figure shows that there are some shared variants used by the two groups (e.g. *affricate, canonical*), but with different weightings, and there are some categories which occur only in the Aboriginal English sample. It is interesting to note that the mainstream Australian English speakers produced over 50% affricates, yet the highest proportion of use of any one variant by the Aboriginal English speakers was canonical /t/ (26%). Aspirated canonical /t/ is used far less by the mainstream Australian English group (16%). While we did not specifically analyse the effect of prosodic position in this paper, it is clearly guiding some of the variation; neither glottal nor ejective tokens were observed intervocalically, while *tap* was only observed intervocalically (but not the sole variant here).

We performed a chi-square test of association to assess whether the distribution frequencies were statistically significant. Overall, the test showed that the two distributions of frequencies across varieties were significantly different, $\chi 2\ (10) = 361.658$, $p < 0.001$. In order to check for differences between the two groups for each of the variants, we used z-tests of proportions, and Bonferroni corrected p-values for multiple comparisons. This showed that the groups differ significantly (at the .05 level, adjusted) for each variant. Despite some shared usage of features (i.e. both groups use *canonical*, both use *affricate*, and so on) the proportions of use of each variant are clearly different.

3.1.1. Aboriginal English: female / male differences

In order to examine possible gendered patterning in the Aboriginal English group, we fitted a multinomial regression, with *canonical stops* as the reference level, and *speaker gender* as the predictor variable. Overall chi-square showed good fit compared to an intercept-only model (as well as a model which attempted to account for context as a constraining factor), $\chi^2(9) = 105.663$, $p < 0.001$. This analysis shows that women are more likely to produce affricates (Wald's $\chi^2 = 12.342$, $p < 0.001$) and men are more likely to produce ejectives, taps, and unreleased glottals (Wald's $\chi^2 = 0.640\text{-}1.61$, $p < .05$). No significant differences were found between men and women for the other variants. Re-fitting the model with *voiced* as the reference level ($\chi^2(9) = 105.663$, $p < 0.001$) shows women are also more likely than men to produce canonical stops (Wald's $\chi^2 = 961.299048$, $p < 0.001$), with the men using other variants as described.

3.1.2. Mainstream Australian English: female / male

In previous work [9], we saw no statistically significant gendered patterns in the variants of post-vocalic /t/ used, unlike what we have described for the Aboriginal English group. Instead we saw an overall preference for what we call "breathy" tokens (i.e. Figure 1), especially affricates and fricatives. In [9], we looked in closer detail at the affricate category given its prevalence, and found sociophonetic conditioning within it, especially "hyper-lengthening" of the release phase which was used significantly more by young speakers and female speakers.

4. Discussion and Conclusion

This study compared variability in post-vocalic /t/ in Aboriginal and mainstream Australian English spoken in a small western Victorian community. Clear sociophonetic patterning of variation across the speaker groups was observed, as well as some variability *within* groups. L1 Aboriginal English speakers use a wider array of post-vocalic /t/ variants than the mainstream Australian English group. The Aboriginal English speakers also used more variants which we can describe as having an overall "glottal" quality (i.e. [ʔt̰] [ʔ] [t']), while the mainstream Australian English speakers, as previously noted, preferred "breathy" variants – e.g. [tˢ] [t̪] [t̪ʰ]. While Aboriginal English speakers used canonical stops frequently, and affricates to some degree, impressionistically it appears that the quality of these is different across the groups. As already noted, the mainstream Australian English speakers hyper-lengthen affricate releases [9], and impressionistically the burst in canonical stops also seems longer overall; this needs independent verification in future work.

That the mainstream Australian English-speaking group use few "glottal" tokens differs from recent phonetic work on Australian English [4], and could be attributable to an urban/regional difference which would be interesting to explore further. The study also showed variability *within* the Aboriginal English sample between males and females (and within the mainstream group with respect to affricate characteristics, as discussed in 3.1.2). Despite varietal differences, where we can say that the Aboriginal English group use fewer "breathy" stops, male Aboriginal English speakers nevertheless used significantly more "glottal" tokens than the female speakers amongst the group; this was not observed in the mainstream Australian English group. The presence of both glottal stops and ejectives is an interesting addition to our knowledge of variability in Englishes in Australia. Where ejectives are concerned, their presence supports the assertion that such variants could occur in Englishes because of "different temporal alignment of the glottal and articulatory components of (pre-)glottalised plosives" [14: 201]; we think a similar interpretation may be true for the full glottal stops as well albeit with supralaryngeal stricture being lost. The findings relating to the presence of ejectives in the data support work predicting a connection between voice quality and consonant type [15,16,17]. While researchers find a connection, however,

they do not claim a one-to-one relationship. Work on Scottish English also shows a strong influence of social and stylistic factors with respect to the use of "glottal" variants (especially ejectives) [17,18]. Our study did not look specifically at age-related variability, but that will likely be another cause for difference within the sample; i.e. it was present for the mainstream sample in [9], and was observed in [4], as well as [18] for Scottish English ejective stops.

Given the /t/ categories used by the Aboriginal English speaking group have a) been previously documented in other Englishes, and b) are not in most cases at all similar to Indigenous Australian languages, results do not suggest a substrate transfer (nor had we predicted that they would). Instead, we interpret the findings as evidence of different production strategies being used across the Aboriginal and mainstream Australian English groups, with clear sociophonetic distribution. While there is overlap in the types of stops used by the two groups, the proportions are vastly different - again, similar to findings for Aboriginal English in [7], and similar to [18] for Scottish English social groups.

One of our main findings is that the distribution of "glottal" vs. "breathy" tokens across the two speaker groups is sociophonetically stratified, and we suggest there is a potential link to voice quality differences. The idea that the realisation of consonants can be linked to voice quality is not unique, although it is certainly under-explored (especially in Englishes in Australia). [15:85] citing [16], explain that "segmental contrasts can provide a testbed for developing the study of voice quality", and [16] proposes a glottal constriction continuum model, which is useful for interpreting our results. We suggest that the mainstream Australian English speakers analysed in this study are on average using a more open glottis, contributing to what we term "breathier" stop variants, and that the Aboriginal English group are likely using a more closed glottis, leading to a greater frequency of glottal or glottalized variants. This idea also supports the opinion that Aboriginal English has a different voice quality to the mainstream variety [6], and is an interesting finding with respect to the overall sociophonetic patterning of post-vocalic /t/ in two varieties of the same language. In this paper we have been deliberately broad in our use of the terms "glottal" and "breathy" and in future work we will acoustically analyse voice quality (H1-H2, H1-A1, spectral tilt) [i.e. 15,16,17], so that we may better understand the distributions observed, as well as confirm possible correlations between type of supralaryngeal articulation and glottal setting.

Although our study has used different analysis techniques, we can align our findings with previous work on Aboriginal English [5,6,7] which shows a maintenance of phonological contrasts used by the mainstream Australian English speaking community, but a difference in phonetic strategies used to implement the contrasts. We have also found that even in L1 Aboriginal English, the phonemic boundary for /t/ is more "porous" [5:134] than in the mainstream variety. The variability that is reported so far in this study and others [5,6,7] indicates that L1 and L2 Aboriginal English speakers rely less on the standard as a target, are exposed to greater variability in input, and in turn use greater variability in production. While this study focused largely on distributions of /t/ variants, the auditory impression of the stops used by Aboriginal and mainstream Australian English speakers is on balance also markedly different. Ejectives are quite emphatic (to varying degrees), full glottal stops, and voiced stops sound relatively marked in an Australian English variety, and fricative variants add a marked breathy quality. As such, we can say that realisation of /t/ is a contributor to a difference in the overall auditory impression of L1 Aboriginal English and mainstream Australian English spoken in Warrnambool.

In summary, this study highlights previously undocumented variation across L1 Aboriginal English and mainstream Australian English. Results suggest clear sociophonetic patterning of /t/ across the two varieties, for females and males within the Aboriginal English group, and for mainstream Australian English females and males with respect to affricate length. We suspect that these patterns may be partly attributable to voice quality and glottal timing differences, which are also potentially sociophonetically determined. Future work will also more closely analyse prosodic position, as well as age-related patterning within the Aboriginal English group.

5. References

[1] Tait, C. & Tabain, M., "Patterns of gender variation in the speech of primary school-aged children in Australian English", in C. Carignan & M.D. Tyler [Eds.], Proceedings of the 16th SST, Paramatta, 65-68, 2016.

[2] Loakes, D. & McDougall, K., "Individual variation in the frication of voiceless plosives", AJL, 30(2):155-181, 2010.

[3] Jones, M.J. & McDougall, K. "The acoustic character of fricated /t/ in Australian English: A comparison with /s/ and /ʃ/", JIPA, 39(3); 265-289, 2009.

[4] Penney, J., Cox, F., Miles, K. & Palethorpe, S., "Glottalisation as a cue to coda consonant voicing in Australian English", JPhon, 66:161-184, 2018.

[5] Malcolm, I. "Australian creoles and Aboriginal English: phonetics and phonology", in K. Burridge & B. Kortmann [Eds.], Varieties of English (3), Mouton, 656-670, 2008.

[6] Butcher, A. "Linguistic aspects of Australian Aboriginal English", Clinical Linguistics & Phonetics 22(8):635-642, 2008.

[7] Mailhammer, R. S. Sherwood and H. Stoakes "Is English on Croker Island an atypical case of Aboriginal English? Evidence from the phonetics of stop consonants", English World-Wide (submitted).

[8] Foulkes, P., Docherty, G.J. & Jones, M.J. "Analysing stops", in Di Paolo, M., Yaeger-Dror,M, [Ed.] Sociophonetics: A Student's Guide, Routledge, 58-71, 2010.

[9] McDougall, K. & Loakes, D. "Australian English /t/ production in Warrnambool: acoustic and sociophonetic variability", Poster presented at *SST-2012*.

[10] Boersma, P. & Weenink, D. (2018). Praat: doing phonetics by computer [Computer program]. Version 6.0.40, retrieved 11 May 2018 from http://www.praat.org/.

[11] Machač, P. & Skarnitzl. R. Principles of Phonetic Segmentation, Epocha, 2019.

[12] Cox, F. & Fletcher, J. Australian English Pronunciation and Transcription, CUP, 2017.

[13] Hajek, J. & Stevens, M., "On the acoustic characterization of ejective stops in Waima'a", INTERSPEECH, 2889-2892, 2005.

[14] Simpson, A. "Ejectives in English and German", in C. Celata & S. Calamai [Eds.] Advances in Sociophonetics, John Benjamins, 189-204, 2014.

[15] Keating, P. & Esposito, C. "Linguistic voice quality", UCLA Working Papers in Phonetics. No. 105, 85-91, 2007.

[16] Gordon, M. & Ladefoged, P. "Phonation types: A cross-linguistic overview", JPhon, 29: 303-406, 2001.

[17] Gordeeva, O. & Scobbie, J. "A phonetically versatile contrast: Pulmonic and glottalic voicelessness in Scottish English obstruents and voice quality", JIPA, 43(3), 249-271, 2013.

[18] McCarthy, O. & Stuart-Smith, J. "Ejectives in Scottish English: A social perspective", JIPA, 43(3), 273-298, 2013.

Interpretations of Uptalk in Australian English: Low confidence, unfinished speech, and variability within and between listeners

Elise Tobin[1], Titia Benders[12]

[1]Department of Linguistics, Macquarie University, Sydney, Australia
[2]ARC Centre of Excellence in Cognition and its Disorders

elise.tobin@mq.edu.au, titia.benders@mq.edu.au

Abstract

This study aimed to determine Australian English (AusE) listeners' interpretations of uptalk, produced by a female AusE speaker. A rating task compared interpretations of uptalk and falling contour utterances. The results indicated that uptalk is perceived to convey lower confidence, reduced emphasis and clarity, and unfinished speech compared to falling contours. Female listeners provided lower Finality ratings for uptalk, compared to male listeners. Confidence and Finality uptalk ratings were variable within listeners. Results are discussed in light of varying listeners' interpretations, and the external and listener-internal factors that may impact uptalk interpretations.
Index Terms: speech perception, intonation, high rising tunes

1. Introduction

Uptalk refers to a declarative utterance produced with a phrase-final rising pitch, which is perceptually similar to the interrogative intonation of yes/no questions present in English varieties [1, 2]. Research on Australian English (AusE) uptalk has predominantly focussed on the phonetic realisation of this intonation contour in production, identifying that uptalk is produced as a L*H-H% tune, according to ToBI transcription conventions [1, 2]. In AusE, these utterances will generally exhibit a 40% pitch increase from the onset of the rise [1]. Uptalk, despite being produced by both male and female speakers [3, 4, 5, 6], is frequently stigmatised by the media as a bad speech habit, and associated with low intelligence and female speech [7].

Perceptual studies of uptalk in AusE have primarily asked whether listeners identify this rising contour as a question or statement [2, 8]. Listeners generally perceive rising contours as statements if the utterance starts with a low pitch accent (i.e. L*H-H%) [2], but the perception of rises is influenced by both the speaker and the listener gender [8].

Few studies on AusE have investigated the influence that a rising intonation may have on listeners' interpretations beyond the question/statement dichotomy. One AusE sociolinguistic study indicated that listeners associate uptalk utterances with uncertainty, deference, youthfulness and low job-suitability [9]. However, it is unclear which variables influenced listeners' interpretations, as the stimulus utterances differed in intonation as well as content. In addition, it is possible that the results were negatively biased by the stigmatic meaning scales chosen by the researchers, which may not reflect current Australian perceptions of uptalk. In an effort to overcome this researcher bias, one American English (AmE) study asked listeners to provide their own interpretations of uptalk; finding that AmE listeners perceive utterances with rising intonation contours to be unfinished, emphasised, clearer and happier [7]. In the present study, we aim to address the stimulus limitations in the AusE study [9], reduce researcher bias by using four of the rating scales from [7], and thus provide an updated understanding of listener interpretations of uptalk in AusE. The second aim of the present study was to address the impact of listener and speaker characteristics on these interpretations, as reviewed in what follows.

Firstly, AmE listeners' interpretations of a speaker producing uptalk are qualified by gender effects. For example, female listeners rated speakers' uptalk utterances to have higher confidence and to be more finished, compared to male listeners' ratings of uptalk [10]. While uptalk has been frequently associated with speaker uncertainty and unfinished speech [7, 9, 10], evidence suggests there is an interaction between gender and intonation interpretation across English varieties [8, 10].

Secondly, listeners' perceptions of uptalk are influenced by listeners' preconceptions about a speaker's characteristics. For example, the implementation of the intonational rise in New Zealand English (NZE) uptalk depends on the social group, and NZE listeners appear to interpret intonation contours based on their assumptions of a speaker's social characteristics, as established through those sociophonetic cues [11]. Listeners' interpretations of uptalk can also be influenced by explicitly provided information about a speaker. For example, word-monitoring reaction times were only affected by phrase-final rising intonation when listeners were informed that the speaker was a 'non-expert', as opposed to when listeners were informed that a speaker was an 'expert' [12]. Such findings suggest that a listener's assumptions of a speaker's mental state influences their interpretation of a spoken utterance [12].

The first question of interest in the present study was whether listeners' interpretations about a female AusE speaker differ, when presented with a falling intonation contour versus a rising intonation contour (i.e. uptalk). The second question asked whether listeners' interpretations of an AusE speaker producing uptalk are influenced by the provision of information that is stereotypically associated or not-associated with the use of uptalk (i.e. information implying that the speaker is, respectively, unintelligent or intelligent). Finally, the effects of listener gender on the interpretations of uptalk were explored.

In line with previous findings [7, 9, 10], it was hypothesised that listeners would perceive a speaker to have less confidence and less finished speech when they produced uptalk compared to a falling intonation. Based on [10], it was also predicted that uptalk would be rated as expressing more emphasis and clarity than a falling intonation. Based on [12], it was hypothesized that the provision of stereotypical 'unintelligent' information would reduce listener ratings of the speaker's confidence and clarity, particularly for statements produced with uptalk.

2. Method

2.1. Participants

A total of 29 Australian born, Australian English speakers between the ages of 18-35 participated (M= 15, F= 14). Participants were recruited in Sydney and were reimbursed with course credit, where applicable. Of these 29 participants, 4 participants were recruited through advertisements posted on the digital learning environment of two units offered by the Macquarie University Department of Linguistics.

2.2. Stimuli

The stimuli were 27 descriptions of iconic Australian concepts, such as 'vegemite' or 'wombat', (e.g. "This is a black food spread made from yeast" or "This is a stocky marsupial", respectively). All descriptions were recorded with a phrase-final rising and a phrase-final falling intonation, by a 25-year-old female native Australian English speaker, who was selected as she produced both uptalk and falling intonations in spontaneous speech. The speaker has received all-round undergraduate training in phonetics, is currently undertaking her PhD in segmental phonetics, and has previously provided spoken stimuli for a range of perception experiments. In order to reduce confounding variables, the speaker produced a similar intonation contour for the first half of the utterance for each rising/falling pair, within the constraints of her natural prosody. In order to obtain natural-sounding stimuli with the desired intonation contours and voice quality, each description was recorded at least 3 times, for both rising and falling intonations.

All stimuli were recorded with an AKG C535EB Condenser Microphone onto an iMac using Presonus Studio Live 16.2.4 AT Mixer in a sound treated studio. Stimuli were recorded at 44.1 KHz, amplitude-normalized, truncated to common temporal landmarks, and digitized as 16 bit WAV files.

Stimuli recordings with a rising intonation were inspected by the first author to confirm consistency with the typical Australian English realisation of uptalk, including the average 40% pitch increase from the onset of the rise. Stimuli recordings with a falling intonation were inspected to confirm the absence of creak, based on auditory impression and striations in the spectrogram. Measures of Fundamental Frequency (*F0*) in Hertz (Hz) were obtained to ensure that the average peak *F0* of the first vowel in each rising/falling pair of stimuli did not differ by more than 20Hz. Durational measures were carried out to ensure each rising/falling pair of stimuli did not differ by more than 0.35 seconds.

Stimuli were presented in the context of a written stereotypical or non-stereotypical vignette (see Table 1), which aimed to elicit specific perceptions about the speaker with fictional information.

Table 1. *Summary of fictional speaker information*

	Stereotypical	Non-stereotypical
Education	Bachelor of Arts	Bachelor of Science
Occupation	Sales assistant	Lab assistant
Spending habits	Make-up	Books

2.3. Procedure

Auditory stimuli were distributed across two lists, with every description only appearing in either falling or rising intonation on a list, and an evenly distributed number of falling/ rising stimuli, for a total of 54 stimuli in each list. The stereotypicality conditions were combined with each of the two stimulus lists, for a total of four experimental groups. Participants were randomly allocated to experimental groups, with an even distribution of male and female listeners across the stereotypical condition (M=8, F=7) and non-stereotypical condition (M=7, F=7).

Participants were seated in front of a laptop computer in a quiet room, and were presented with auditory stimuli via Sennheiser 380 Pro headphones at a self-selected comfortable listening volume. The listening task involved a practice phase and a test phase. During the practice phase, the rating scales and four examples of the auditory stimuli were presented, and participants had the opportunity to clarify questions with the experimenter. The vignettes were provided at the start of the test phase and immediately following the two enforced 10-second breaks during the test phase.

Participants were instructed to rate the speaker's utterances based on their first impression. Participants rated each stimulus in regards to Confidence, Emphasis, Clarity and Finality (see Table 2), on four continuous scales from 'Strongly Disagree' (corresponding to a score of 0) to 'Strongly Agree' (corresponding to a score of 100). Stimuli were presented one at a time; participants had the option of replaying each stimulus and they were able to change their ratings before clicking to continue to the next stimulus. Participants completed two exit questionnaires at the conclusion of the listening task. The first questionnaire was a self-assessment of task performance, and was not further analysed for the present study. The second questionnaire provided a manipulation check of the two conditions, to assess whether the vignettes had elicited the desired impressions of the speaker; these responses are briefly mentioned in the discussion.

Table 2. *Impression Rating scales*

Rating Scale	Rating prompt
Confidence	The speaker is confident.
Emphasis	The speaker is emphasizing her point.
Clarity	The speaker is speaking clearly
Finality	The speaker is finished talking.

2.4. Statistical analysis

The confirmatory analysed the participant Mean (M) scores for each rating scale and contour as dependent variables in 2x2x2 mixed-effects Anova's assessing the effects of contour type (rising vs falling, within-subjects), gender (male vs female, between-subjects) and condition (stereotypical vs non-stereotypical, between-subjects) on listeners' ratings in each scale. To maintain an Anova-wise alpha level of 0.05, we adopted an alpha level of 0.007 which was Bonferroni corrected for the seven comparisons within each Anova.

Upon inspection of the raw data, it was observed that individual participants' rating scores deviated more from their mean for rising compared to falling contours, particularly for the Confidence and Finality rating scale. To explore this within-participant variability, each participants' standard deviation was divided by their mean score to obtain a coefficient of

variance (SD/M) and analysed, using the previously mentioned 2x2x2 Anova. All statistical analyses were conducted in the *R* statistical programming environment [13].

3. Results

Research question 1 asked whether listeners' perceptions about a speaker differ, when presented with a falling vs. rising intonation contour. Research question 2 asked whether listeners' judgements are influenced by the provision of stereotypical or non-stereotypical speaker information. While participants' free responses in the exit questionnaire stated that the speaker information provided was insufficient to form judgements about the speaker, their ratings on the exit questionnaire were consistent with the vignettes. This suggests that the stereotypicality manipulation was successful and may have influenced participants' responses.

As predicted for Research Question 1, the speaker was rated as more confident when producing a falling (M= 85, SD= 8.5) compared to a rising contour (M=62, SD=15.5, F(1, 25)= 39.5, p<0.0001; see Figure 1). She was also rated as showing more emphasis when producing a falling (M=77, SD=13.7) compared to a rising contour (M=63, SD=14.9, F(1, 25)= 13, p<0.0014). Contrary to the expectations, the speaker was perceived to speak more clearly when producing a falling (M=84, SD=10.8) compared to a rising contour (M=79, SD=10.7, F(1, 25)=15, p<0.0007). In contrast with the hypotheses for Research Question 2, there was no interaction between condition and contour for the Confidence, Emphasis or Clarity rating scales. Moreover, there were no other significant two- or three-way interactions.

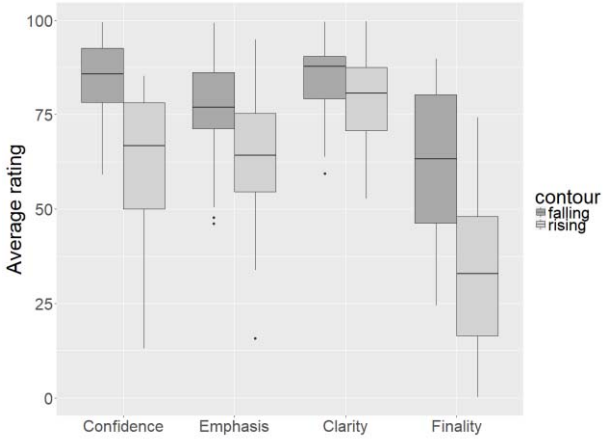


Figure 1: *Average ratings for falling vs rising contours*

As predicted for Research Question 1, falling contours were rated as being more finished (M=63, SD=26.8) than rising contours (M=34, SD=20, F(1, 25)=70.3, p<0.0001). This main effect is qualified by a significant interaction between contour and gender for the Finality rating (F(1, 25)=14.4, p<0.0008; see Figure 2). This interaction revealed that male listeners' Finality ratings were more similar for falling (M=60, SD=25) and rising contours (M=42, SD=22), compared to female listeners' Finality ratings for falling (M=64, SD=28) and rising contours (M=25, SD=17).

The interaction between condition and contour for the finality rating was marginally significant (F(1, 25)=4.4, p<0.045), suggesting a larger difference in Finality ratings between falling and rising contours in the non-stereotypical, compared to the stereotypical condition. However, this result was not interpreted as strong evidence, as the Bonferroni corrected alpha-value was 0.007.

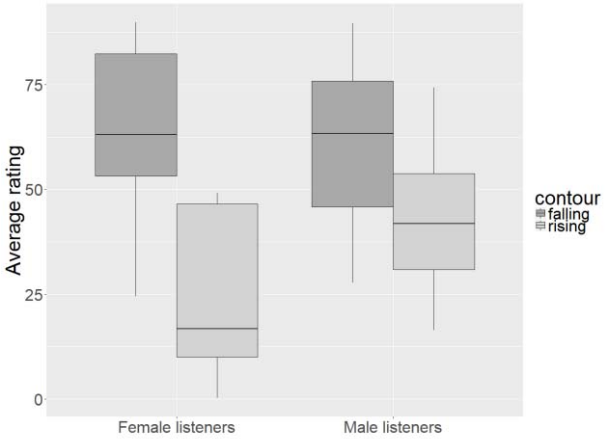


Figure 2: *Contour and gender interaction for finality rating*

Exploratory analyses were conducted on the SD/M dependent variable to determine whether within-participant variability was larger for rising compared to falling contours. A significant effect of contour was found for the Confidence rating (F(1, 25)=31.5, p<0.0007) as well as the Finality rating (F(1, 25)=22.1, p<0.0007), with participants' Confidence and Finality ratings being more varied for rising compared to falling contours (see also Figure 3). A marginally significant effect of contour was found for the Emphasis rating (F(1, 25)=4.5, p<0.044), but this result was not significant following the Bonferroni correction.

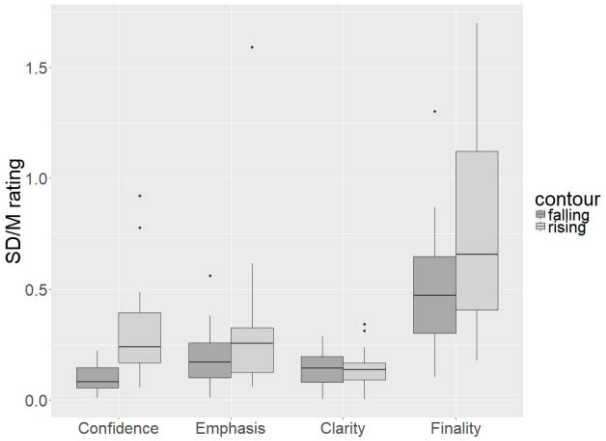


Figure 3: *SD/M of ratings for falling vs rising contours*

4. Discussion

This study aimed to provide an updated understanding of listeners' interpretations of a female Australian English (AusE) speaker producing uptalk. Previous American English (AmE) studies have found that listeners perceive uptalk utterances as less confident, unfinished, and with more emphasis and speech clarity, compared to utterances with a falling intonation [7]. Consistent with previous perception studies across English varieties [7, 9, 10], AusE listeners in the present study associated uptalk utterances with lower confidence and unfinished speech. However, they provided lower emphasis and clarity ratings for the uptalk utterances.

Because we did not expect listeners to rate uptalk utterances to have reduced emphasis and clarity, the present findings

suggest that interpretations of uptalk may differ across English varieties. However, one limitation of the present study is that the scales used may not reflect AusE listeners' uptalk interpretations. For example, L2 English speakers have reported to interpret uptalk as ironic or sarcastic, depending on the situation and their knowledge of the speaker [14]. Future research would benefit from first surveying AusE speakers on the meanings that they associate with uptalk. The unexpected findings regarding Emphasis and Clarity ratings may also be ascribed to the stimuli being produced by a phonetically trained speaker rather than being spontaneously produced. The speaker's phonetic and linguistic knowledge may have influenced her speech in a recording situation. For example, the speaker may have deliberately expressed similar Emphasis for falling/rising stimuli and her lab-speech enunciation may have removed any differences in Clarity. Future research would benefit from collecting listener judgements of spontaneous uptalk productions. Further research with a range of speakers would also allow for increased generalizability of listeners' perceptions of uptalk.

The present study built on findings that NZE and AmE listeners' perceptions of uptalk can be affected by inferred or provided speaker characteristics [11, 12]. In the present study, we provided participants with either a stereotypical or non-stereotypical vignette, in an attempt to elicit negative or positive perceptions about the female speaker's intelligence, but did not find any significant effect thereof on listener ratings. Future research assessing the influence of contextual information should incorporate audiovisual stimuli, thus providing implicit information via communicative cues that are absent from the auditory signal, for every utterance in the stimulus set. It is also possible that the lexical features and content of the stimuli affected listener ratings, resulting in a non-significant effect. For example, the stimuli descriptions were factually correct, which may have increased positive perceptions about the speaker's intelligence.

Previous research had also identified that listener gender interacts with the interpretation of intonation [8, 10]. In contrast to previous work with AmE listeners [10], the present study did not find that female listeners associate uptalk with a higher confidence than males and even found that female AusE listeners have a stronger association between uptalk and unfinished speech compared to male listeners. These results suggest that there are different ways in which male and female listeners interpret uptalk across English varieties [8, 9, 10, 11].

The observed gender effect on AusE listeners' interpretations of uptalk raises the question of gender effects in uptalk production. Specifically, the weaker association between uptalk and unfinished speech for male listeners may reflect a higher ratio of uptalk usage by male, compared to female AusE speakers. Conversely, the gender of the speaker influences listener perceptions of intonation rises in AusE [8]. Future research would need to determine whether the observed uptalk interpretations and interactions with listener gender are specific to a female AusE speaker.

In addition to the between-listener differences associated with gender, there was also significant variation within participants for the Confidence and Finality ratings for uptalk compared to falling intonations. Between participant variability in ratings has been previously observed [7, 9, 10], but to the best of our knowledge, this study is the first to observe that individuals vary more in their own response to individual uptalk utterances compared to utterances with a falling intonation. This implies that listeners' interpretations of uptalk cannot be solely explained in terms of external variables, such as intonation, context, and speaker gender. Future research is needed to identify the factors that influence utterance-to-utterance fluctuations in listeners' interpretations of uptalk.

In summary, this study has found that AusE listeners perceive uptalk to sound less confident and less finished compared to a falling intonation. Female listeners rated uptalk utterances as less finished compared to male listeners, and participants were more variable in their Confidence and Finality ratings for uptalk compared to a falling intonation. In light of previous research, these results suggest that listeners' interpretations of uptalk differ due to external and listener-internal factors, and fluctuate from one utterance to the next.

5. Acknowledgements

We thank Louise Ratko for providing the speech stimuli used in this study. Thanks to Macquarie University Phonetics Lab for helpful discussions and feedback on methodology and data analysis, in particular Anita Szakay and Julien Millasseau.

6. References

[1] J. Fletcher and D. Loakes, "Patterns of rising and falling in Australian English," in *Proc. of the 11th Australian Int. Conf. on Speech Sci. & Technology*, *December 6-8,2006,* P. Warren, C. Watson, Eds. Auckland, New Zealand, 2006, pp.42-47.

[2] J. McGregor and S. Palethorpe, "High rising tunes in Australian English: The communicative function of L*and H*pitch accent onsets," *Aust. J. Linguist.,* vol. 28, pp.171–193, 2008.

[3] J. Fletcher, E. Grabe, and P. Warren, "Intonational Variation in Four Dialects of English: the High Rising Tune," in *Prosodic Typology: The Phonology of Intonation and Phrasing*. Oxford University Press, 2005. [Online] Available: Oxford Scholarship Online.

[4] K. Allan, (1984). "The Component Functions of the High Rise Terminal Contour in Australian Declarative Sentences," *Aust. J. Linguist.*, vol.4, pp.19–32,1984.

[5] E. Levon, "Gender, interaction and intonational variation: The discourse functions of High Rising Terminals in London," *J. Sociolinguistics*, vol.20, pp.133–163, 2016.

[6] T.J. Linneman, "Gender in jeopardy: Intonation Variation on a Television Game Show," *Gender & Society*, vol.27, pp.82–105, 2013.

[7] J.C. Tyler, "Expanding and Mapping the Indexical Field: Rising Pitch, the Uptalk Stereotype, and Perceptual Variation," *Journal of English Linguistics,* vol.43, pp.284–310, 2015.

[8] E. Schmidt, B. Post, C. Kung, I. Yuen, and K. Demuth, "The effect of listener and speaker gender on the perception of rises in AusE," *in Proc. of the 18th Int. Congress of Phonetic Sciences (ICPhS2015)*, Glasgow, UK, 2015.

[9] G.R. Guy and J. Vonwiller, "The meaning of an intonation in Australian English," *Aust. J. Linguist.,* vol. 4, pp.1-17, 1984.

[10] V. Shokeir, "Evidence for the stable use of uptalk in South Ontario English," *U. Penn Working Papers in Linguistics,* vol. 142, 2008.

[11] P. Warren, "The interpretation of prosodic variability in the context of accompanying sociophonetic cues," *Lab. Phonology: J. Assoc. Lab. Phonology*, vol.8, pp.1–21, 2017.

[12] J.M. Tomlinson and J.E Fox Tree, "Listeners' comprehension of uptalk in spontaneous speech," *Cognition,* vol.119, pp. 58–69, 2011.

[13] R Core Team (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

[14] E. Tobin, "Prosody in Australian English: Second language learners' perceptions of Uptalk and its impact on production and comprehension", Unpublished manuscript, Macquarie University, Sydney, NSW, Australia, 2018.

Coronal stop VOT in Australian English: Lebanese Australians and mainstream Australian English speakers

Josh Clothier[1, 2] *and Debbie Loakes*[1, 2]

The University of Melbourne, ARC Centre of Excellence for the Dynamics of Language

joshuajc@unimelb.edu.au; dloakes@unimelb.edu.au

Abstract

This paper analyses variability in voice onset time (VOT) in Australian English (AusE) coronal stops, as produced by those of mainstream, and Lebanese heritage ethnicities. Data are derived from the AusTalk corpus, and from a new corpus of Lebanese Australian ("LAus") speech in Melbourne comprising 30 speakers aged 18—30. Both groups use short- and long-lag VOT for /d/ and /t/ respectively; however, speakers with Lebanese heritage exhibit substantially more pre-voiced /d/ tokens. Group-wise, there are fine phonetic differences between groups. Lebanese Australian individuals exploit a greater range of VOT values but there is no evidence of transfer of Lebanese Arabic VOT system. Instead, speakers deploy coronal stop VOT to index ethnic identities as Lebanese Australians.

Index Terms: sociophonetics, Australian English, ethnicity, plosives, voice onset time, acoustic phonetics

1. Introduction

Early in the 21st century, Cox [1] argued that Australian English would see greater variability along ethnocultural lines, in parallel with the greater ethnolinguistic diversity of Australian society. Cognisant of this, we pursue this study of ethnolectal variability, now well into the second decade of the 21st century. Further, the coronal plosive /t/ in AusE has been shown to carry substantial sociolinguistic information. In non-initial contexts, it has been shown to reduce, with, for example, spirantised realisations associated with female speech and tapped realisations associated with male speech [2]. Furthermore, Horvath [3] argued that a highly aspirated/affricated realisations of /t/ ([t^{S}]) was associated with gender (women) and ethnicity (speakers with Greek heritage) in her study of the sociolects of Sydney. The percept of "heavy aspiration" described by Horvath, would likely be associated with particularly long-lag VOT durations.

VOT—the period between the release of a plosive's closure and the onset of voicing of the following segment—has been shown to vary as a function of system internal and external factors. In English, VOT provides a somewhat reliable cue to the phonological feature [voice]; while there are complications, the feature [+voice] is associated with short-lag VOT, and [-voice] is realised with long-lag VOT. [4] has shown that the height of the following vowel can affect VOT duration, with stops preceding high vowels tending to longer VOT durations compared to stops preceding non-high vowels. System-externally, VOT has been shown to vary as a function of sex and gender and cross-linguistic evidence [5] shows that a gender-based (i.e., socially learnt and constructed) argument for women's longer VOT in English, which is a reliable effect across different studies, is sounder.

Labov [6] has argued that ethnolectal features, with respect to ethnolectal "substrates" (speakers' heritage languages, which may or may not be spoken by speakers themselves) operate in different ways depending on variety, language, social context, and the specific variable under question. Clyne et al [7] argue for "stabilised transference" in the development of ethnolectal features, and this has been shown to be evident in the speech of Lebanese Australians, wherein temporal differences in the structure of VC rhymes have been identified in Sydney [8].

While it has been shown that bilingual speakers are able to maintain distinct VOT systems for each of their languages, including specifically in the case of children bilingual in Arabic and English [9], given that substrate effects *may* be observed in speakers of ethnocultural varieties, this should be investigated as a possible factor. Broadly speaking, the stop-voicing system of Arabic retains phonetic voicing in VOICED stops, with VOICELESS stops realised with short-lag VOT. This is in contrast (again, broadly speaking) with English VOT, which is classified as an aspirating language, wherein VOICELESS stops use long-lag VOT and VOICED stops use short-lag VOT (with pre-voicing occurring in some prosodic contexts and for some speakers).

1.1. Aims & research questions

In this *sociophonetic* study, we investigate the patterning of AusE coronal stop (/t d/) VOT between groups of speakers with "mainstream" and Lebanese ethnicities. Specifically, we ask:

1. Does coronal stop VOT vary as a function of *ethnicity* between these two ethnic groups?
2. Do the two *ethnicity* groups deploy the same VOT durational strategies for making VOICING contrasts in AusE?
3. How does gender interact with ethnicity in the realisation of /t d/ stop VOICING?
4. How does the height of the following vowel affect the realisation of VOT in these stops for these speakers?

2. Method

2.1. Speakers

Data come from two sources: a corpus collected by the first author (the "M-LAusE Corpus") and the AusTalk corpus [10]. 20 speakers were selected from the AusTalk corpus who satisfied the following criteria:

- Were raised (and acquired AusE) in the Melbourne region.
- Were aged 18-30 at the time of recording
- In responding to questionnaire items regarding cultural heritage, did *not* provide information suggesting their classification as "mainstream" Australian speakers would be inappropriate (i.e., heritage is Anglo-Celtic Australian; no

information about parents or grandparents from a non-majority ethnicity)

Speakers in the M-LAusE corpus (N = 30) are age matched to the young-adult group of the AusTalk corpus (i.e., 18-30y) and were also raised and acquired AusE in Melbourne but have parents and/or grandparents who were born in Lebanon. For these speakers, home language prior to starting school is either *English only*, *English and Lebanese-Arabic,* or *Lebanese-Arabic only.* This variability is inherent in the sampling: it reflects the variability in the community. It should be noted that the study (and the larger project) analyses the effects of ethnic identity, not of bilingualism, which have been explained elsewhere. For thorough accounts of the effects of bilingualism on VOT, see, e.g., [9, 11].

2.2. Materials

2.2.1. Phonetic data

Data used in this study were elicited using the AusTalk [10] wordlist task. Within the wordlist, there are seven words with initial /d/ and eight with initial /t/ in stressed syllables. For the M-LAusE corpus, data were recorded in a sound treated recording studio at the University of Melbourne using Charter Oak E700 Microphones with Aphex 1100 MKii mic preamp and a BSS DPR-402 compressor. All data were recorded onto a Macbook Pro running Bootcamp Windows 7 through a Digidesign 003 rack firewire soundcard using Samplitude Pro X Suite as the software, and were recorded at 16 bit 44.1 KHz. Details of the AusTalk recording protocols are available in [10]. AusTalk data used in this study come from the Melbourne subset, which were recorded in the same sound treated recording booth as the M-LAusE corpus.

2.3. Analysis

2.3.1. Segmentation

Segmentation was aided by the AutoVOT [12] supervised learning algorithm, which estimates boundaries for VOT based on 7 acoustic parameters described in [13]. We trained four models for use in this analysis: one for each variety and VOICING subset. Training used a total of 200 pairs of sound files and Praat TextGrids from the AusTalk corpus, and 225 pairs of sound files and TextGrids from the M-LAusE corpus. Pseudo random selection of tokens from the corpora was employed to ensure sufficient representability across speakers and places of articulation[1]. Training TextGrids were segmented with boundaries at the onset of the burst and the zero-crossing preceding the first upward oscillation into semi-periodic waveform associated with vocalic energy. Predicted segments of VOT were visually inspected and corrected as necessary, based on principles used in segmentation of training tokens.

2.3.2. Statistical analysis

Statistical models were built in lme4 [14] using R [15] in the RStudio IDE [16]. After building a maximally specified lmer model, the step function in the lmerTest [17] package was used to aid identification of the best fitting model (1) below. This model was confirmed using likelihood ratio tests, χ^2 (12) = 212.63, $p < .001$.

[1] This study is part of a larger study which looks at VOT across all initial stops, /p t k b d g/

The model which best fits these data, and from which we derive the post-hoc contrasts discussed below is:

$$vot \sim voicing + ethnicity + gender + V\_height + (1 \mid spkr\_id) + (1 \mid word) + voicing{:}ethnicity + voicing{:}gender + ethnicity{:}gender + voicing{:}V\_height + V\_height + gender{:}V\_height + voicing{:}ethnicity{:}gender + ethnicity{:}gender{:}V\_height \quad (1)$$

To test the effect of home language, mixed effects logistic regressions were fit, with gender and home language as fixed predictors and speaker and word random intercepts. Model comparison shows the best fitting model has significant main effects of gender, χ^2 (1) = 14.299, $p = 0.000156$.

3. Results

Figures 1-3 in this section plot density distributions, as is typical in the contemporary VOT literature [e.g., 18]. Vertical lines have been added at the 0 ms mark to indicate the cut-off between negative (pre-voiced) and positive (pre-voiced) tokens.

As described in model (1) above and shown in Figure 1 below, VOICING is a significant predictor of VOT duration in these data, as expected; as a group, speakers use VOT duration to make the English VOICING contrast in word initial /t d/. As a group, Lebanese Australians' VOICED stops have significantly longer short-lag VOT (M = 22.5 ms) compared to the mainstream group (M = 18.7 ms), $p = 0.0085$.

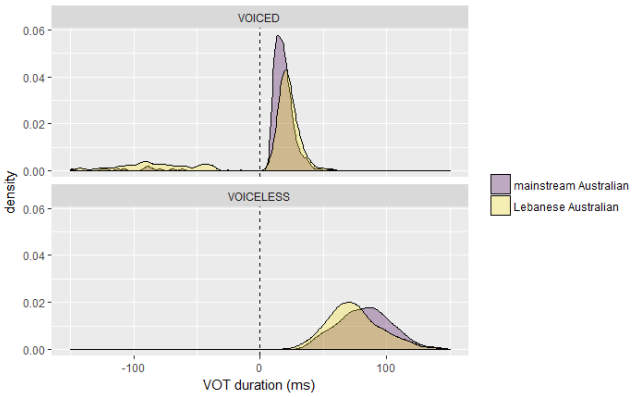


Figure 1. *Distributions of VOT for /t, d/ as a function of ethnic heritage.*

The Lebanese Australian group also has substantially more pre-voiced tokens of VOICED stops (N = 150) compared to the mainstream group (N = 12). VOICELESS stops' VOT are shorter (M = 75.7 ms) for the Lebanese Australians compared to the mainstream group (M = 82.8 ms); however, this difference is not significant (p = .0786). Specific sociophonetic patterns are discussed for the VOICED (/d/) and VOICELESS (/t/) stops below.

3.1. Gender

There was a significant main effect of gender, with women tending to longer positive VOT overall. Post-hoc analysis shows that this difference only retains significance for the Lebanese heritage group (p = 0), and not the mainstream group (p = 1). While the interaction term is retained as contributing to the model of best fit described above in (1), post-hoc analysis

with Bonferroni adjusted *p*-values shows no significant interaction between these factors. However, some interesting trends are evident in the distributions, when we consider the entire VOT continuum. Notably, Lebanese Australian men are responsible for the majority of the pre-voiced tokens of /d/ across the two corpora.

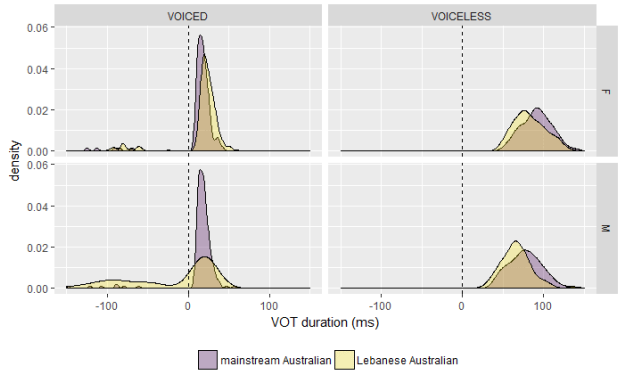


Figure 2. *Distributions of VOT for /t, d/ as a function of ethnicity and gender*

Table 1 shows frequencies of pre-voiced tokens across speaker groups and genders, demonstrating that the clear majority of these tokens are produced by men with Lebanese heritage. An analysis at the speaker level shows that across the 5 (of 20) speakers with mainstream ethnicity who pre-voice, the most pre-voiced tokens produced are by two speakers who produce four pre-voiced tokens each. Within the Lebanese heritage group, though, there are 20 speakers who produce pre-voiced tokens of /d/, ranging from between just one and 19 (i.e., a majority of this speaker's /d/ tokens). The distribution of/t d/ VOT durations for the speaker producing the most pre-voiced /d/, LM-008, is shown below in Figure 3. While he does produce 19 tokens of /d/ with pre-voicing (negative VOT duration) and just two with short-lag VOT (positive VOT duration), his VOICELESS stops, /t/, lie within the long-lag region (positive VOT duration).

Table 1. *Counts of prevoiced /d/ tokens by group and gender*

Group	Gender	Pre-voiced /d/ tokens
Mainstream	F	5
	M	7
Lebanese Australian	F	24
	M	126

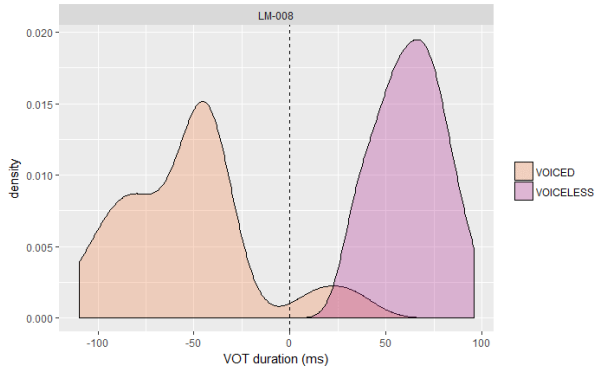


Figure 3. *Distribution of coronal stop VOT for individual speaker LM-008 who uses pre-voicing in /d/*

3.2. Vowel height

There was no significant main effect of the height of the following vowel, $p = .1143$. Analysis of post-hoc effects shows multiple interactions. Notably, /t/ when followed by high vowels is overall longer than when followed by low vowels ($p = .0192$; see figure 4.). There is also a significant three-way interaction between ethnicity, gender, and vowel height, wherein there is a most marked difference between the two ethnic groups for male speakers in stops preceding a high vowel; here, the Lebanese Australian males' have the shortest VOTs using this three-way comparison, $p < .05$.

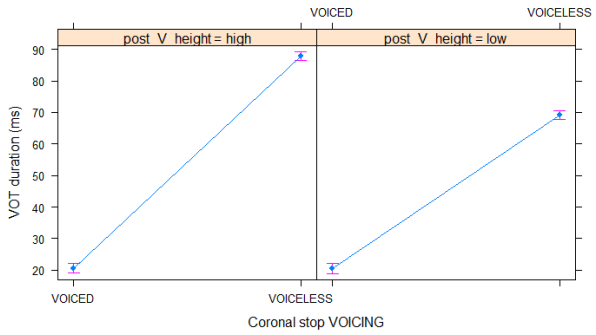


Figure 4. *Interaction between vowel height and coronal stop voicing: effect on VOT*

4. Discussion

4.1. Ethnicity

4.1.1. VOT durations

We asked whether coronal stop VOT varied as a function of ethnic heritage between these two groups, and the results show that there is a clear effect of ethnic heritage on VOT patterning in AusE. The finding is suggestive of the emergence of the kind of ethnolectal variability suggested for AusE in the 21st century by Cox [1] and suggests that speakers are using VOT to do identity work. This aspect is explored in detail in the larger study, which provides a rich, quantitative sociophonetic account of ethnic identity.

4.1.2. VOICING contrasts

We also asked whether the stop VOICING contrast was realised in the same ways in both groups of speakers and, on a group basis, they are. AusE speakers with mainstream and Lebanese ethnicities tend to use the "classic" English short-lag/long-lag VOT distinction for the phonological categories [+voice] and [-voice] respectively. There are, however, some deviations from this pattern at the individual level. There is evidence of more speakers with Lebanese heritage using pre-voicing in VOICED stops /d/, which may be suggestive of a transfer effect from Arabic. This is not a wholesale, systemic effect on the VOT system, though, as the VOICELESS stops for these speakers are still realised with long lag VOT; furthermore, as a group, the Lebanese Australians studied here use longer durations of VOT in the short-lag category for /d/ than the mainstream AusE group do. This could be explained by a variable grammar in which competing forces between the transfer effect of Arabic, which might be pulling VOT durations downwards (into pre-voicing for /d/ and towards short-lag for /t/), is mediated by competition from the sociolinguistic valence of /t/ in AusE [cf 19] (and we must remember that these speakers *are* speaking

AusE and are embedded within larger Australian social contexts), which has been shown elsewhere to be heavily aspirated, particularly by non-mainstream speakers [2, 3].

4.2. Gender

We have shown that women in this study, in accordance with other studies on English varieties elsewhere in the world [e.g., 20], have longer positive VOTs than men do. The present analysis does not elucidate the existing understanding of why this might be the case. We can only concur with others [e.g., 5] in that this is a language-specific effect; our understanding of this as also being a social effect is enhanced by the finding which is significant for the analysis of the corpora together, for the M-LAusE corpus, but *not* the mainstream AusE corpus.

4.3. Vowel height

Our findings of a vowel height effect accord with other literature [e.g., 4], wherein VOT durations for /t/ are longer before high vowels. Interestingly, the inclusion of this variable draws out more differences between the two ethnic groups along the gender dimension, suggesting subtly different ways of speaking AusE for men and women; that is, on this factor, Lebanese Australian women behave more like the mainstream, while Lebanese Australian men diverge further, again, strengthening sociological explanations for the study's findings.

5. Conclusions

We have identified some key features of coronal stop VOT in AusE, including between-group differences on the basis of ethnicity, and gender. There are VOT differences suggestive of specifically gendered ways of speaking AusE for Lebanese Australians: in terms of pre-voicing (Lebanese men use the most pre-voicing), VOT durations (all women's VOT durations are longer, but Lebanese women's are longest), and the effect of vowel height (all /t/ VOTs are longer when followed by a high vowel, except for Lebanese men's). While these findings are robust, they benefit from further elucidation, which is ongoing in the larger study, where we find further effects of sociological factors, such as social networks and religious identity, but continue to find no effect of home language, or reported proficiency in Lebanese Arabic. In other words, the sociophonetic effects on stop realisation are stronger than linguistic effects. Analysis of the VOT system of the full stop series /p t k b d g/ will provide a clearer picture of how VOT is operating in AusE, and will also help to answer some of the questions that this paper raises about the potential for transfer of features from Lebanese Arabic (i.e., would we see the same pattern of pre-voicing and aspiration in the labial series given that there is no contrast at this place in Lebanese Arabic?). We provide here a first account of VOT from the Melbourne AusTalk corpus, and an account of VOT showing clear ethnocultural patterns in AusE coronal stops, suggestive of the kind of ethnocultural variability predicted by Cox.

6. Acknowledgements

This research is supported by an Australian Government Research Training Program (RTP) Scholarship. I am grateful to all those who have provided constructive feedback on this work.

7. References

[1] F. Cox, "Australian English pronunciation into the 21st century," *Prospect,* vol. 21, no. 1, pp. 3-21, 2006.
[2] L. Tollfree, "Variation and change in Australian English consonants," *Blair & Collins (eds.),* pp. 45-67, 2001.
[3] B. M. Horvath, *Variation in Australian English: The sociolects of Sydney* (Cambridge Studies in Linguistics, no. 45). Cambridge: Cambridge University Press, 1985.
[4] J. Berry and M. Moyle, "Covariation among vowel height effects on acoustic measures," *J Acoust Soc Am,* vol. 130, no. 5, pp. EL365-71, Nov 2011.
[5] F. Li, "The effect of speakers' sex on voice onset time in Mandarin stops," *The Journal of the Acoustical Society of America,* vol. 133, no. 2, pp. EL142-EL147, 2013.
[6] W. Labov, "Mysteries of the Substrate," in *Social Lives in Language – Sociolinguistics and multilingual speech communities: Celebrating the work of Gillian Sankoff*, M. Meyerhoff and N. Nagy, Eds.: John Benjamins, 2008, pp. 315-326.
[7] M. Clyne, E. Eisikovits, and L. Tollfree, "Ethnic varieties of Australian English," in *English in Australia*, vol. 26, 2001, pp. 223–238.
[8] F. Cox and S. Palethorpe, "Timing differences in the VC rhyme of standard Australian English and Lebanese Australian English," *Proceedings of the 17th International Congress of Phonetic Sciences,* pp. 528-531, 2011.
[9] G. Khattab, "VOT production in English and Arabic bilingual and monolingual children," in *Leeds Working Papers in Linguistics*, vol. 8, D. Nelson and P. Foulkes, Eds., 2000.
[10] D. Estival, S. Cassidy, F. Cox, and D. Burnham, "AusTalk: an audio-visual corpus of Australian English," presented at the 9th Language Resources and Evaluation Conference, Reykjavik, Iceland, 2014. Available: https://austalk.edu.au/sites/default/files/AusTalk-LREC2014-final.pdf
[11] M. Antoniou, C. T. Best, M. D. Tyler, and C. Kroos, "Language context elicits native-like stop voicing in early bilinguals' productions in both L1 and L2," *J Phon,* vol. 38, no. 4, pp. 640-653, Oct 2010.
[12] J. Keshet, M. Sonderegger, and T. Knowles, "AutoVOT: A tool for automatic measurement of voice onset time using discriminative structured prediction ", 0.91 ed, 2014.
[13] M. Sonderegger and J. Keshet, "Automatic measurement of voice onset time using discriminative structured prediction," *J Acoust Soc Am,* vol. 132, no. 6, pp. 3965-79, Dec 2012.
[14] 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.
[15] Microsoft and R Core Team, "Microsoft R Open," 3.4.4 ed, 2018.
[16] R. Team, "RStudio: Integrated Development Environment for R," 1.1.447 ed. Boston, MA: RStudio, Inc., 2016.
[17] A. Kuznetsova, P. B. Brockhoff, and R. H. B. Christensen, "lmerTest: Tests in Linear Mixed Effects Models," 2.0-30 ed, 2016.
[18] G. J. Docherty, D. Watt, C. Llamas, D. Hall, and J. Nycz, "Variation in voice onset time along the Scottish-English border," in *Proceedings of the 17th International Congress of Phonetic Sciences*, 2011, pp. 591-594.
[19] L. Newlin-Łukowicz, "From interference to transfer in language contact: Variation in voice onset time," *Language Variation and Change,* vol. 26, no. 3, pp. 359-385, 2014.
[20] S. P. Whiteside, L. Henry, and R. Dobbin, "Sex differences in voice onset time: A developmental study of phonetic context effects in British English," *The Journal of the Acoustical Society of America,* vol. 116, no. 2, pp. 1179-1183, 2004.

Preliminary Investigations into Sound Change in Auckland.

Catherine Watson[1], Brooke Ross[2], Elaine Ballard[1], Helen Charters[1], Richard Arnold[2] and Miriam Meyerhoff[2]

[1]University of Auckland
[2]Victoria University Wellington

Abstract

New Zealand English is traditionally characterised by raised /e/ and /æ/ vowels and the retraction of /ɪ/, with very little regional variation. In the following study we report findings from a study in Auckland that suggests otherwise. 32 young speakers (aged between 18-25yrs) from three suburbs in Auckland were recorded reading a passage. A further 6 older speakers (between 45-70yrs) were also recorded. Hand-corrected formants from the recorded vowels with sentence stress were analysed. The results showed there that sound change has occurred between the younger and older speakers, with the /e/ and /æ/ vowels lowering and backing, and /ɪ/ lowering. This lowered /e/ and /æ/ is noticeable when compared with the results to other studies with Modern NZE speakers recorded outside Auckland. The implication of these results is discussed.

Index Terms: New Zealand English, Auckland, Vowels, Sound Change

1. Introduction

New Zealand English (NZE) is traditionally characterised by raised /e/ and /æ/ vowels and the retraction of /ɪ/ vowel (e.g. [1]). Recent research indicates that the raising of /e/ has encroached on the acoustic space of the /i:/ vowel resulting in the diphthongisation of the latter vowel [2, 3]. It has been assumed that this is the case for all NZE speakers. According to Gordon & Maclagan [4] regional differences are not strongly marked in NZE for the front vowels. Bauer [5] also states "In New Zealand you can take seven hours to drive from Wellington to Auckland, and not be able to "hear any difference in the English that is spoken when you arrive."

This study presents an acoustic analysis looking at phonetic diversity in Auckland, New Zealand's largest city, and scrutinises these assertions. Auckland has been New Zealand's largest city for over 100 years, and yet there has never been a major linguistic analysis of the English spoken there [6]. This is a serious oversight, given the important role NZE has played in the developments into the theories of dialect contact, leveling and new dialect formation [7]. Furthermore, in the last 10 years there has been increased migration into New Zealand, Auckland specifically. In 35% of Auckland's suburbs, no ethnic group is more than 50% of the population; many speakers were born elsewhere and many more have grown up using different varieties of English as the spoken norm. Given that Cheshire and colleagues [8, 9] found extreme ethnic mixing and recent migration can result in leveling and changes in speech in urban centres such as London , we ask whether this is the case for the NZE accent spoken in Auckland.

2. Investigation 1

2.1. Speakers

As part of the Auckland Voices project [6], 33 speakers from three suburbs in Auckland (Mt Roskill n= 14 (8 women, 6 men), Papatoetoe, n=13 (6 women, 7 men) Titirangi, n=6 (3 women, 3 men)) were recorded reading a 390 word passage [10]. The participants were aged between 18 and 25 years. Our speakers were either New Zealand born or arrived in the country under the age of seven. The three suburbs were deliberately chosen: Titirangi is a predominantly Pākehā (NZ European) community; Papatoetoe is a well-established ethnically mixed community and Mount Roskill is a community undergoing demographic change

2.2. Data Preparation.

Speakers' recordings were digitised and transcribed in ELAN [11]. WebMAUS (NZ English option) [12] was used for the automatic phonetic labelling, then phonetic boundaries and labels were checked and hand corrected where necessary in PRAAT [13]. PRAAT text grids were converted into the EMU-webApp [14] format for formant calculation and analysis. Formant tracks were checked and hand corrected when necessary, then vowel targets were marked as per [15]. Finally, the first and second formants were extracted at the vowel targets marked according to the criteria in [15]). The remainder of the formant analysis was done in R [16] using EmuR [17]). Only vowels that carried phrase stressed were studied in this analysis. Over 2800 monophthongs were analysed. Table 1 gives the number of tokens per vowel.

Table 1: *The number of vowel tokens per suburb group.*

	iː	ɪ	e	æ	ʌ	ɑː	ɒ	ɔː	ʊ	uː	ɜː
Mt Roskill	227	87	167	150	116	84	119	118	14	44	25
Papatoetoe	218	90	166	140	112	78	108	109	20	56	22
Titirangi	100	51	77	86	57	41	52	52	15	19	13

2.3. Data Transformation

The distribution for men and women Bark scaled formant plots are shown in Figure 1. It is possible to make the frequency observations directly comparable by making a simple linear transformation. If the formant values for a woman speaker are (F_1;F_2) then we can convert them into the values closer to values (M_1;M_2) typical for a man speaker by the relation

$$\begin{bmatrix} F_1 \\ F_2 \end{bmatrix} = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} + \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \begin{bmatrix} M_1 \\ M_2 \end{bmatrix} \quad (1)$$

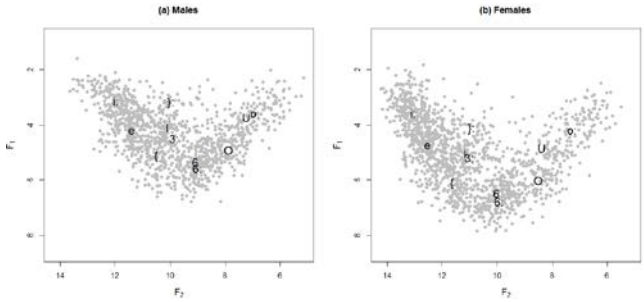


Figure 1: *Raw formant observations on the Bark scale for men and women separately.*

where $\mu_1, \mu_2, A_{11}; A_{12}; A_{21}; A_{22}$ are a set of 6 numerical constants. The calibration of the relationship requires the choice of 3 points in the men's distribution and the corresponding 3 points in the women's distribution onto which they will be mapped

$$\left(\begin{bmatrix}M_{11}\\M_{12}\end{bmatrix},\begin{bmatrix}F_{11}\\F_{12}\end{bmatrix}\right),\left(\begin{bmatrix}M_{21}\\M_{22}\end{bmatrix},\begin{bmatrix}F_{21}\\F_{22}\end{bmatrix}\right),\left(\begin{bmatrix}M_{31}\\M_{32}\end{bmatrix},\begin{bmatrix}F_{31}\\F_{32}\end{bmatrix}\right)$$

where M_{ij}, and F_{ij} are the centroids of the point vowels /i:. a:, ɔ:/. Using these values, we can then solve (1) to find the six unknown constants. Once we have those values we can then use equation (1) to transform the men's data onto the women's speaker space. It amounts to a simple translation, rotation and stretching of one plane so that the three anchoring points match. Figure 2 shows the effect of the transformation on the vowel space for men and women when the men's vowel data is transformed to match the women's. The sample mean locations of the anchoring vowels /i:. a:, ɔ:/ are matched exactly.

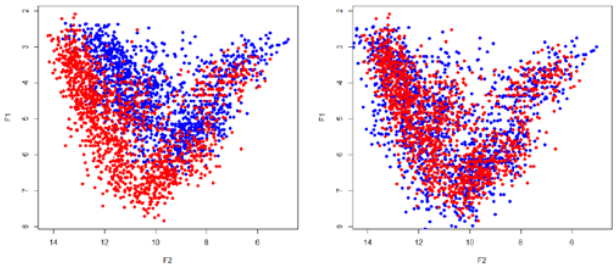


Figure 2 *Raw men's (Blue) and women's (Red) data points (left), Transformed men's (blue) vowel space on to women's (red) vowel space. (right*)

Figure 3 shows the effect of the transformation on the vowel centroids for men and women when the men's vowels are transformed to match the women. The anchoring vowels /i:, a:, o:/ are matched exactly.

2.4. Results of Investigation 1

Usual analysis of F1 and F2 treats them separately, and consequently separate tests for changes in each formant are performed. Here we model F1 and F2 values simultaneously in the 2-D formant space. This allows for increased statistical power in which changes in any direction in the plane can be detected. For each vowel we built four linear mixed effects models for observations of the first and second formant. Each model had speaker as a random effect and the fixed effects for each model are given in Table 2. All statistical analyses were performed in R [16], the linear mixed effect models were calculated using the `lme()` function in the `nlme` package and the comparison between models was done via ANOVAs using the `anova()` function from the `nlme` package. Note due to the small numbers of /ʊ/ per speaker (typically 2, see Table 1), the results for this vowel will not be reported in the statistical analysis. Note also the reporting will be done phonetically, not in terms of actual F1 and F2 values. For full details readers are referred to [18]

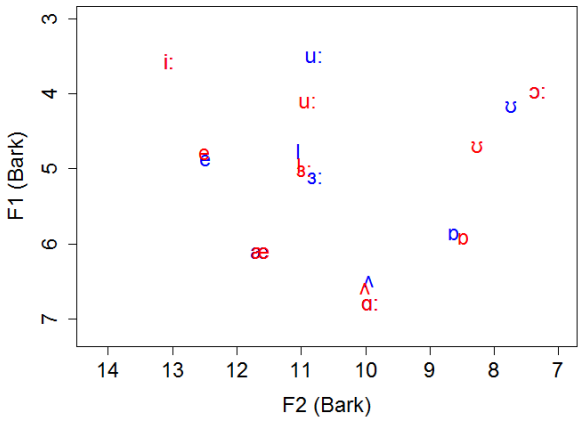


Figure 3: *Vowel centroids after the transformation. women's centroids are plotted in red, and men's centroids in blue.*

Table 2: *The* four Linear *models in the investigation.*

Model Name	Fixed effects	Random effects	Observation
null	Type	Speaker	Formant value
g1	Type*Suburb	Speaker	Formant value
g2	Type*Sex	Speaker	Formant value
g3	Type*Sex*Suburb	Speaker	Formant value

Table 3: Significant *differences when null model is compared to model with fixed effects of type and suburb (g1).*

	Degrees of Freedom	AIC Difference (null – g1)	Log Likelihood Ratio	P value
/i:/	12	4.926	12.926	0.04
/ɜ:/	12	1.9943	9.994	0.012

When the null model was compared with g1 (Type*Suburb) there were only two vowels where the models differed significantly, these were /i:/ and /ɜ:/, see Table 3. Further analysis using t-tests and setting Mt Roskill as the reference suburb revealed that for /i:/ Mt Roskill was significantly higher and fronter in the vowel space than for Papatoetoe (subP:t(31)=2.81 p<0.01; typeF2:subP:t(1054)=-2.55,p<0.05) , Mt Roskill is also significantly fronter than Titirangi ((typeF2:subT:t(1054)=-2.55,p<0.05) For /ɜ:/ Mt Roskill is significantly higher than Titirangi (subT:t(30)=2.35,p<0.05) , and significantly fronter than Papatoetoe (typeF2:subP: t(85)=-2.21,p<0.05). These differences can be seen in Figure 4, although note for space reasons a specific plot showing only suburb differences has not be included in this paper.

When the null model was compared to g2 (Type*Sex) there were only significant differences between men and women for one vowel, which was /u:/ (see Table 4). Post hoc t-test revealed that the /u:/ for the women speakers was significantly lower than that for the men speakers (sexM:t(32)=-3.88,p<0.001). These differences can be seen in Figure 3.

Table 4 S*ignificant differences when the null model is compared to the model with fixed effect of type and sex.*

	Degrees of Freedom	AIC Difference (null-g2)	Log Likelihood Ratio	P value
/u:/	10	8.0195	12.01354	0.0025

In the final comparison the null mode was compared to g3 (Type*Suburb*Sex). There were significant results for the five vowels /i: ɪ æ u: ɔ:/, Table 5 lists the statistics for the significant results. The random effects estimation for the g3 model is given in Figure 4. The large coloured symbols show the centroids for each group, and the coloured lines connect the centroids for the individuals within a group to the group centroid.

Table 5 S*ignificant differences when the null model is compared to the model with fixed effects of type, suburb and sex*

	Degrees of Freedom	AIC Difference (null-g3)	Log Likelihood Ratio	P value
/i:/	18	0.146	20.146	0.0279
/ɪ/	18	8.6464	28.446	0.0015
/æ/	18	2.301	22.3	0.0136
/u:/	18	6.6485	26.649	0.003
/ ɔ:/	18	0.4216	20.421	0.0255

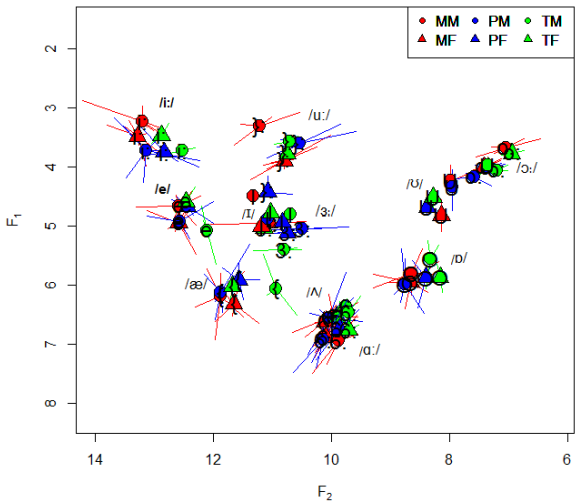


Figure 4: *Random effect estimation of F1 and F2 for all utterances of all 11 vowels, from the three suburbs (Mt Roskill (red), Papatoetoe (blue) and Titirangi (green)) and for both men (circle) and women (triangle) speakers.*

In the post-hoc t-tests, Mt Roskill is the reference suburb, and women is the reference gender. The post-hoc analysis revealed the following. For /i:/ the Mt Roskill women are significantly fronter than the Papatoetoe women (typeF2:subP: t(1051)=-2.24, p<0.05). For /ɪ/ Mt Roskill women are significantly lower than the Mt Roskill men (sexM: t(28)=-4.51, p<0.001) Papatoetoe men (subP:sexM:t(28)=2.64, p<0.05) and Titirangi men (subT:sexM t(28)=2.63, p<0.05) The Mt Roskill women /ɪ/ is also significantly fronter than the Titirangi men (typeF2:subT:sexM: t(417)=-2.71,p.<0.05). For /æ/ the Mt Roskill women are significantly fronter than the Titirangi men. (typeF2:subT:sexM: t(713)=-2.68, p<0.05). For /u:/ the Mt Roskill women are significantly higher than the Papatoetoe women (subP: t(28)=3.07,p<0.001). Further Mt Roskill women are significantly lower and backer than the Mt Roskill men (sexM: t(28)=3.08, p<0.001; typeF2:sexM: t(199)=2.30 ,p<0.05). Finally for /ɔ:/ the Mt Roskill women are significantly lower than the Mt Roskill men (sexM:t(28)=-3.08,p<0.01), and significantly higher than the Papatoetoe and Titirangi men (subPsexM:t(28)=3.13,p<0.01; subTsexM:t(28)=3.17,p<0.01).

3. Investigation 2

3.1. Speaker and Data preparation details.

In this second investigation we compare the young speakers in the Auckland Voices corpus to the older speakers. At the time of writing only one set of older speakers had been labelled – 6 female speakers from Titirangi. These speakers were aged between 45 and 70 (mean age 59 years, sd. 9 years). We compared these older speakers to all the young women speakers in the Auckland Voices corpus. All the latter group were considered because they were very few differences between the these speakers in the suburb analysis, only /i:/ and /u:/ between the Mt Roskill and Papatoetoe (see section 2.4). The data preparation was prepared as outlined in Section 2.2. Since all speakers were women no data transformation was necessary.

3.2. Results

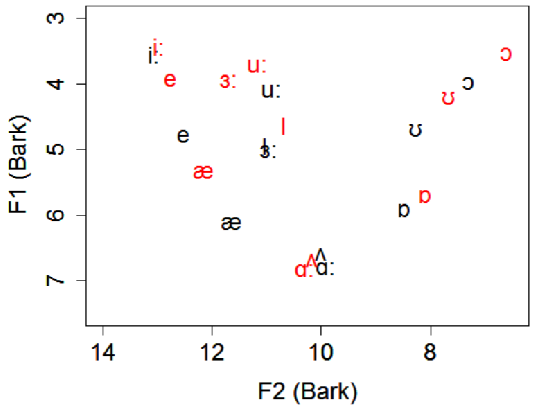


Figure 5: *F1/F2 plots of the young Female speakers (black) and older female Speakers (red) in the Auckland Voices corpus.*

As in Section 2.4 we compared linear mixed models. Each model had speaker as a random effect, the fixed effects for the null model was the formant type (F1 and F2) and the comparison model (g4) was formant type interacting with age (Y and O). Once again, the models were compared using ANOVAs. In the post-hoc t-tests the young women were the reference group. Post hoc analysis revealed the following. For /ɪ/ the younger speakers were significantly lower than the older speakers (ageO: t(23)= -2.806, p<0.01). For /e/ younger speakers were lower and backer than the older speakers (ageO: t(23)= -6.510,p<0.000; typeF2ageO:t(580)= 5.965,p<0.0001).

The same was found for /æ/, (ageO: t(23)= -6.00,p<0.0001; typeF2ageO:t(506)= 7.791,p<0.0001). For /u:/ the younger speakers were significantly backer than the older speakers (typeF2:ageO: t(168)= 2.679,p<0.01). Finally for /ɔ:/ the younger speakers were fronter and lower than the older speakers (ageO: t(23)= -4.686,p<0.001; typeF2ageO:t(370)= -2.091,p<0.05). These differences can be seen in Figure 5. It was not possible to create a figure like Figure 4 for this data because there were not enough tokens for /ʊ/ and /ɜ:/ to create models, therefore only the centroids have been used.

Table 6: S*ignificant differences when the null model is compared to the model with fixed effects of type and age*

	Degrees of Freedom	AIC Difference (null – g4)	Log Likelihood Ratio	P value
/ɪ/	10	6.26	10.266	<.01
/e/	10	21.18	25.184	<.0001
/æ/	10	26.25	30.348	<.0001
/u:/	10	2.71	6.9113	<.01
/ ɔ:/	10	12.39	16.396	<.001

4. Discussion and Conclusion

Two comparisons were performed in this study. In the first we compared the speech from young speakers in three different suburbs of Auckland. For / e ɑ: ʌ ɒ/ there were no differences. However, there were clear suburb differences for /i: ɪ æ ɔ:/. Of note is the finding that Mt Roskill women differed more than any other single group. Given that Mt Roskill is the suburb with biggest demographic changes in the last couple of decades, it is perhaps not surprising that the young women in this population look to be the most innovative speakers. That aside, we do need to increase the number of speakers in our Titirangi data set. This is the group that we would expect to be the most conservative, but we had less than half the number of speakers in this group than the other two groups. Our analysis to date has focused on the vowels from reading passages; but as noted these passages did not provide enough /ʊ/ tokens and the number of tokens for /u:/ and /ɜ:/ were also very low. To gain a fuller picture of all the NZE vowels we will also need to analyse the vowels from interviews recorded with the same participants.

In the second study we compared the speech of the young and old women. There were significant differences for all three of NZE's signature vowels /ɪ e æ/. Notably /e æ/ have dropped and retracted over time. Interestingly the vowel space from the older women speakers is more what we would have expected in NZE [e.g. 3, 15, 19]. The /ɪ/ movement is also very interesting; it has fronted, but not risen. There are two preliminary findings from this study. The first finding suggests that there has been a sound change in Auckland English. At this stage, only the older women from Titirangi have been analysed, so the next step is to include the speech from the older speakers in the other two suburbs, as well as looking at both women and men, to see how consistent this change is. The second finding is perhaps even more profound. It suggests that there is a difference between the English spoken in Auckland and elsewhere in New Zealand in spite of earlier observations (e.g. [4, 5]). In contrast to the present study, two very recent studies [3,19] have also maintained the idea of uniformity within NZE by reporting that young speakers of NZE produce high /e/, very near the /i:/ vowel. The speakers from these studies are notably not from Auckland (from the Hamilton and Wellington region in the one study and Hamilton and Christchurch region in the other study).

5. Acknowledgements

We thank the NZ Royal Society Marsden fund for supporting the project, the participants in the Auckland Voices project, our research assistants, and the anonymous reviewers.

6. References

[1] Watson, C.I., Maclagan, M. & Harrington, J. (2000). Acoustic evidence for vowel change in New Zealand English. Language Variation & Change, 12: 51-68.

[2] Maclagan, M. & Hay, J. (2007). Getting fed up with our feet: contrast maintenance and the New Zealand English "short" front vowel shift. Language Variation & Change, 19: 1-25.

[3] Warren, P. (2017). Quality and quantity in New Zealand English vowel contrasts. Journal of the International Phonetic. Association, First View, 1-26. doi.org/10.1017/S0025100317000329

[4] Gordon, E. & Maclagan, M. (2004). Regional and social differences in New Zealand: Phonology. In Kortmann, B. et al. (eds.), A Handbook of Varieties of English: Volume 3. Berlin and New York: Mouton de Gruyter. pp 64-76.

[5] Bauer, L. (1994). English in New Zealand. In Burchfield, R (Ed.), The Cambridge history of the English language: volume 5. Cambridge: Cambridge University Press. Pp 382-429.

[6] Meyerhoff, M. (2017) "Community dependencies Connections and discontinuities in Auckland City." New Zealand Linguistic Society, Auckland, Nov 23-24, 2017.

[7] Trudgill, P., Gordon, E., Lewis, G. & Maclagan, M. (2000) Determinism in new-dialect formation and the genesis of New Zealand English. Journal of Linguistics, 36(2), 299-318.

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

[9] Cheshire, J., Fox, S., Kerswill, P. & Torgersen, E. (2013) Language contact and language change in the multicultural metropolis. Revue Française de Linguistique Appliquée, 17(2), 63-76.

[10] Holmes, J., Bell, A. & Boyce, M. (1991). Variation and Change in New Zealand English: A Social Dialect Investigation (Project report to the Foundation for Research, Science & Technology). Wellington: Victoria University of Wellington, Linguistics Department.

[11] Sloetjes, H. & Wittenburg, P. (2008). Annotation by category – ELAN and ISO DCR. In: Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008)

[12] Kisler, T. Reichel U. D. & Schiel, F. (2017): Multilingual processing of speech via web services, Computer Speech & Language, 45, 326–347

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

[14] Winkelmann, R., & Raess, G. (2014). Introducing a web application for labeling, visualizing speech and correcting derived speech signals. In the 9th International Conference on Language Resources and Evaluation (LREC 2014), pp. 4129-4133.

[15] Watson, C. I., Harrington, J., & Evans, Z. (1998). An acoustic comparison between New Zealand and Australian English vowels. Australian Journal of Linguistics, 18(2), 185-207.

[16] R Core Team (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/

[17] Winkelmann, R., Jaensch, K., Cassidy, S. & Harrington, J. (2018). emuR: Main Package of the EMU Speech Database Management System R package version 1.0.0

[18] Ross, B (2018) "An Acoustic Analysis of New Zealand English vowels in Auckland" Unpublished Master's thesis submitted to Victoria University, Wellington in fulfilment of the requirements for the degree of Masters of Arts

[19] Watson, C. I., Maclagan, M. A., King, J., Harlow, R., & Keegan, P. J. (2016). Sound change in Māori and the influence of New Zealand English. Journal of the International Phonetic Association, 46(2), 185-218.

The development of cross-accent recognition of familiar words by bilingual and monolingual toddlers: The effect of pre-exposure

Tina Whyte-Ball[1], Catherine Best[1], Karen Mulak[1, 2], and Marina Kalashnikova[1]
[1]MARCS Institute, Western Sydney University
[2] Department of Hearing and Speech Sciences, University of Maryland

Abstract

Bilinguals differ from monolinguals in the development of speech perception [1], word learning [2] and language discrimination [3]. In light of this, we investigated whether early bilingual experience influences the development of cross-accent word recognition. We tested monolingual and bilingual 17-month-olds' recognition of familiar words spoken in an unfamiliar accent in a listening preference task following pre-exposure to a multi-talker story told in their native or the unfamiliar accent. Monolinguals and bilinguals recognised the familiar words in both accents. Taking into account previous findings that without pre-exposure, monolingual 19-month-olds failed on the same task and stimuli [4], the current results imply that hearing the story in either accent supported recognition of familiar words in the unfamiliar accent. Furthermore, the findings suggest that bilingual experience did not significantly affect this ability.

1. Introduction

All speech contains phonetic variability stemming from gender, speech rate, speech style, talker, and accent differences in pronunciation patterns. Adult native perceivers can "hear through" many of these differences to accurately recognise a word, or else can rapidly adapt after brief exposure. Although accented speech may initially affect comprehension and speed of processing, adults need only brief talker exposure to adapt to their accent [5]. Therefore, adults quickly become familiar with the features of a speaker or group of speakers and use this information to successfully comprehend their speech.

For young children, this task is not as easy. Young children are word recognition novices. Prior to 19 months, they fail to reliably recognize familiar words spoken in an unfamiliar accent. However, by 19 months, they appear to have grasped two complementary principles necessary for separating meaningful from non-meaningful phonetic variation [6]. Infants need to differentiate the crucial difference between the type of phonetic variation that signals a lexical distinction (*phonological distinctiveness*) and the type that does not (*phonological constancy*). For example, an adult American English perceiver who recognises that the pronunciation of the word 'mice' in American English (AmE) [maɪs], is the same as the phonetically different Australian English (AusE) pronunciation [mɑes] has grasped *phonological constancy* [6]. On the other hand, an adult AmE perceiver who recognises that the AusE pronunciation [mes] is not 'mice' but the contrasting word 'mess', has grasped *phonological distinctiveness*.

Thus, grasping phonological constancy means in part that infants recognise that the phonetic variation in a word can violate native-accent phonemic boundaries without changing the identity of the word. The perceptual attunement account [6], proposes that for phonological constancy and therefore cross-accent word recognition to emerge, young children must shift their attention at the lexical level from detailed phonetic patterns to higher order phonological abstractions that allow them to recognise when the phonological structure of a word is the same regardless of the phonetic variations. Based on this account, word recognition in non-native accents remains challenging until infants develop the ability to rely on phonologically specified, rather than phonetically detailed word forms.

Early studies on word recognition have provided insight into the emergence of phonologically specified word forms, which is proposed to begin emerging by 14 months, but only if there is contextual support and reduced task demands. By 19 months, attention to phonologically specified word forms becomes more robust [7]. To test when phonological constancy emerges, Best and colleagues [6] presented AmE-learning infants at 15 and 19 months with familiar and unfamiliar words in both their familiar native accent as well as an unfamiliar regional accent to which they had no previous exposure, Jamaican Mesolect English (JaME). Toddlers' propensity to preferentially listen to familiar over unfamiliar words was exploited to index word recognition in a listening preference task. Fifteen-month-olds showed a familiar word preference in the native accent, but showed no preference between familiar and unfamiliar words in the unfamiliar JaME accent. Nineteen-month-olds, however, showed a familiar word preference in both accents, suggesting robust access to phonologically specified word forms, or *phonological constancy* [see also 8].

A later study demonstrated that 15-month-olds acquiring Canadian English were able to recognise words in an unfamiliar accent (AusE), but only after receiving exposure to a story read by their parent at least four times over 2 weeks prior to testing [9]. They tested whether infants' familiarity with the story might facilitate accent adaptation. If this were true, increasing familiarity with words in the story exposure phase would allow infants better access to the words, perhaps leading to recognition when produced in an unfamiliar accent. After hearing the story read by their parents at home, children heard the story read in the unfamiliar accent prior to the listening preference task with known versus nonsense words. Following the story phase, 15-month-olds showed a preference for known over nonsense words. Also, they found that only the infants who heard the story at home listened longer to the known over the nonsense words. This showed that the toddlers at 15 months adapted to the unfamiliar accent after exposure.

To account for how phonological constancy emerges, [10] proposed that accumulating exposure to between- and within-speaker variation fosters the emergence of abstract phonological knowledge of word structure. Drawing from the findings summarised above and the proposed impact of prior exposure, we explore whether infants exposed to high variability in their language environment might have an advantage in development of phonological constancy. Language input to bilinguals is higher in variability than that to monolinguals [11]. Thus, given the proposed role of variability in the development of phonological constancy, we investigated the role of (1) variability in infants' daily input and (2) pre-exposure to an unfamiliar accent on their word-recognition skills in their native

and an unfamiliar accent. For this purpose we tested 17-month-old monolingual and bilingual infants on a listening-preference word recognition task in their native and an unfamiliar accent after receiving pre-exposure to a brief story in the native or the unfamiliar accent. This age group was selected because [12] found that 17-month-olds in the word-preference task differed by their vocabulary size in whether or not they recognized familiar words in their native accent. Furthermore, supporting the effect of vocabulary size at 17 months, [13] linked 17-month bilinguals' inability to identify words in both familiar and unfamiliar accents to their reduced exposure to each language which results in reduced vocabulary size. This age group will not only allow the findings of the current study to be compared with previous findings in [12, 13], but will also allow the current study to show whether bilinguals can benefit from pre-exposure at an age when they had failed to show the emergence of phonological constancy. Finally, testing infants at 17 months could further support findings that younger infants can perform well on certain word recognition tasks if appropriate support is provided [2, 9].

Most research investigating the effects of accent variation on speech processing by infants and young children has focused on monolingual infants. The findings from the study with AmE-learning infants at 15 and 19 months reported above [6] showed that phonological constancy is evident in monolinguals by 19 months but not at 15 months. Later studies considered how productions of individual phonemes in an unfamiliar accent may be perceptually assimilated to phonetic categories in the native accent by adults, and the possible effect this might have on cross-accent word recognition by young children. Perceptual assimilation is the perceivers' detection of phonemes of a non-native accent relative to the native accent. In light of the potential effect of perceptual assimilation on acceptance of phonetic differences, Category-Goodness (CG) and Category-Shifting (CS) type accent differences in consonants and vowels were identified and introduced to familiar word listening preference studies [4, 14]. CG differences are perceptible accent differences in pronunciation of a given phoneme, but the difference does not change the perceived category of the phoneme in the listener's native accent. For example 'spoon' pronounced in AusE [spʉ:n] and in JamE [spu:n], differ in the front vs. back realisation of /u/. It would be expected that despite the phonetic difference between AusE and JaME, the JaME production of /u/ would be heard as /u/ (though perhaps as an odd version) by a native AusE speaker. CS differences, on the other hand, are cross-accent phonetic differences that adults perceive as a phonemic contrast in the native accent, for example, the stressed vowel in 'baby' in AusE [bæ̱ɪbɪ] vs. JaME [be̱bɪ]. An AusE adult would hear the vowel in the JaME pronunciation as a different vowel (<*bee*>) than in the AusE pronunciation (<*bay*>).

Prior studies have found that both CG consonant and vowel differences fail to impair cross-accent word recognition even at 15 months [4, 14]. On the other hand, both 15-month-old and 19-month-old infants failed to recognise familiar words in the unfamiliar accent when the accented words displayed CS vowel differences [14]. Thus, although 19-month-olds can shift their attention from specific phonetic patterns to a more abstract phonological structure when non-native-accented words display CG vowel or consonant variations from the native accent, this ability is hindered when CS vowel differences are involved. Therefore, the current study focused on words that display CS vowel differences between accents, to provide a clear test of whether story pre-exposure would yield adaptation to the unfamiliar accent.

Limited research on accent variation and lexical processing has focused on bilingual children, with the exception of a recent study that compared bilinguals' and monolinguals' cross-accent word identification. Seventeen-month-old bilinguals showed poorer native and cross-accent word identification than monolinguals, but seemed to "catch up" to monolingual peers by 19 months [8, 13]. There are several reasons why bilinguals' developmental trajectory might be different in comparison to monolinguals' based on differences in their language input [11]. First, while total language exposure is not expected to differ between monolinguals and bilinguals, bilinguals are likely to hear less speech in a given language than monolinguals as their exposure is split between two languages. This might lead to a bilingual disadvantage because reduced vocabulary size, which results from reduced exposure to each language [11], may cause a delay in bilinguals' cross-accent word identification [13]. Conversely, though, there might be a bilingual advantage due to other features of their input. Firstly, as opposed to monolinguals who only need to represent the phonological structure of words in one language, bilinguals have to represent the phonological structure of words in two languages. Furthermore, bilinguals must track information in each language to separate and differentiate their languages. Also, there is high variability in their input; bilingual exposure is "noisy" (two languages and accented versions of each). These factors could lead to an advantage in accommodating phonetic variation in accents as bilinguals have more experience with negotiating phonetic variation in their input as compared to monolinguals. A bilingual advantage would also be consistent with the proposal that exposure to variation can be beneficial to developing phonological constancy [10]. Therefore, since increased variability in the input along multiple dimensions appears to help listeners determine which cues are most important in processing speech with differences in pronunciation [10], the variability in bilingual input could possibly be advantageous in cross-accent lexical processing, leading to better word recognition by bilinguals than monolinguals when accented pronunciations display CS vowel differences.

1.1 Aim and Predictions

The aim of this study was to compare bilinguals' and monolinguals' ability to benefit from short-term pre-exposure to accented speech in a word-recognition paradigm. The findings from [9] suggested that exposure to an accent can immediately impact infants' ability to shift from specific phonetic patterns to a more abstract phonological structure to allow them to accept a wider range of pronunciations. Given that children at the age when phonological constancy emerges (around 19 months) fail to generalise to a non-native accent when the words display CS vowel differences, we selected infants at a younger age (17 months) to test whether exposure to talker and accent variability impacts phonological abstraction during this transition period for recognising familiar words that display CS vowel differences when produced in an unfamiliar accent. We predict that if the impact of pre-exposure on phonological abstraction is immediate, 17-month-old toddlers will show a listening preference for familiar words that display category-differing CS vowel differences after exposure. Alternatively, if the impact of pre-exposure on phonological abstraction takes some time to develop, similarly to the infants in [14], 17-month-olds may fail to generalise word recognition to words in an unfamiliar accent that display CS vowel differences even after exposure.

In addition, given that bilinguals receive more variable in-

put than monolinguals, and less exposure to each language (reduced input + reduced vocabulary), this could result in one of two between-group differences:

1. *Bilinguals outperform monolinguals*: More variable input leads bilinguals to accommodate better to variation. If bilinguals benefit from their "noisy" input, pre-exposure to JaME should facilitate word recognition in the JaME accent by bilinguals relative to monolinguals.
2. *Monolinguals outperform bilinguals*: Bilinguals' reduced input in each language may instead lead to delayed development of phonological constancy [12]. We reasoned that if bilinguals are delayed in cross-accent word recognition relative to monolingual peers, pre-exposure to the unfamiliar accent should assist word recognition in JaME more in monolinguals than in bilinguals at this age.

2. Method

2.1 Participants

Thirty-two 16- to 18-month-olds, 16 monolinguals (M_{age} = 16.9 months) and 16 bilinguals (M_{age} = 16.8 months) from Sydney, Australia participated in the study. The monolinguals' exposure to non-native languages or non-AusE accents ranged from zero to no more than four hours per week [10]. Consistent with bilingual lexical development studies [e.g., 15], bilingual participants had to be receiving a maximum of 70% exposure to one language and a minimum of 30% exposure to the other. On average, bilinguals heard each language 35-65% of the time, as measured by parental report on a language exposure questionnaire [13]. The toddlers selected for the bilingual group were children with diverse language backgrounds. They came from homes where AusE English and another language were being spoken regularly. These languages included Cantonese, Arabic, French, Russian, Korean, and Spanish. Although bilinguals were exposed to different languages other than English, their different individual language profiles were not expected to affect performance as shown in studies such as [11]. An additional nine infants were tested but excluded from analyses due to extreme fussiness (n = 6) or failure to meet the monolingual or bilingual language exposure criteria (n = 3).

2.2 Stimulus materials

Pre-exposure consisted of the short children's story "Chicken Little" told in either the children's familiar AusE accent or in the unfamiliar JaME accent. The AusE story was read by three female native speakers of AusE, and the JaME story was read by three female native speakers of JamE. Each speaker read the story in child-directed speech, and the exposure passage was created by combining segments from each speaker such that their voices alternated throughout the pre-exposure phase. There were 11 story segments concatenated across talkers, each accompanied by a different story board illustration. The entire passage lasted four minutes.

The test stimuli consisted of 32 target words that had been used in a previous study on word recognition [12]. The words were produced by the same AusE and JaME talkers who produced the stories for the pre-exposure phase. The target words displayed CS vowel differences between the AusE and JaME accents. The target words were 16 toddler-familiar words that young children are very likely to know, and 16 low-frequency, unfamiliar adult words that they are extremely unlikely to have ever heard (8 single-syllable and 8 two-syllable words per type). The toddler-familiar words were selected on the basis that they appear in most toddlers' early vocabularies and/or in children's books well-known to toddlers. Overall, the words had a mean frequency of 77% in 17-month-olds' expressive vocabularies [16]. The adult-unfamiliar, low-frequency words occur in English adult corpora at frequencies of less than 5 per million words [17]. Multiple tokens of each word were recorded in citation form. The final tokens of each target word (two per speaker per word) were selected based on similarity across accents in voice quality and speech quality.

2.3 Procedure

Infants sat on their parent's lap in a sound-attenuated booth, facing a centrally positioned video monitor. During the procedure, the parent listened to music over headphones to prevent them from inadvertently influencing their infant's behaviour. The study was divided into two phases: a pre-exposure phase and the listening preference phase which included two tests. During the pre-exposure phase, infants listened to either the AusE or JamE story while viewing the storyboard on a central monitor. Half of the children in each group heard the AusE story (8 bilinguals, 8 monolinguals) and the other half heard the JaME story (8 bilinguals, 8 monolinguals).

Participants then completed the two listening preference tests, one in the native AusE accent and one in the unfamiliar JaME accent, with test accent order counterbalanced across children in each group. Each test had eight trials: four trials with toddler-familiar words, and four with adult-unfamiliar words, presented in alternation. On each trial of a given test, words of the designated word set were played in random order with an inter-stimulus interval of 750 ms. Words were presented continuously as long as the child remained fixated on a coloured checkerboard on an LCD monitor (max trial duration = 30 s). The audio ceased when the child looked away, and the trial ended if the child did not look back to the monitor within 2 s. At the end of a trial, the checkerboard flashed until the child's gaze was recaptured, and a new trial began. The listening preference task takes advantage of toddlers' tendency to preferentially listen to familiar over unfamiliar words [18], which was used to index familiar word recognition. For the duration of the experiment, eye gaze direction was monitored and recorded via a hidden video camera focused on the child from below the monitor.

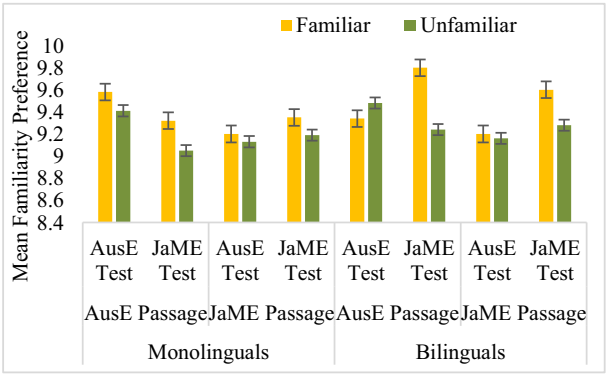


Figure 1: *Mean log-transformed looking time to familiar versus unfamiliar words across test conditions and language experience. Error bars represent standard error of the mean.*

3. Results

A mixed-effects linear model was fit using toddlers' log-transformed looking time with test trial type (unfamiliar vs. familiar words) as the dependent variable, and test accent (JaME vs. AusE) and language experience (bilingual vs. monolingual) as fixed effects, and participant and test order as random effects. A main effect of trial type revealed a preference overall for familiar over unfamiliar words, $F(1,476.85) = 5.33$, $p = .021$ (see Figure 1). Three interactions approached significance: test trial type x test accent, $F(1, 477.02) = 3.46$, $p = .063$, test accent x language experience $F(1, 477.01) = 3.28$, $p = .070$, and test accent x passage accent $F(1, 476.98) = 3.27$, $p = .071$.

4. Discussion

We tested whether pre-exposure to a children's story told by multiple AusE vs. JaME talkers would have different effects on monolinguals' vs. bilinguals' cross-accent word recognition. Results suggest that the impact of pre-exposure on phonological abstraction in cross-accent word recognition is rapid and generalises across words in a child's vocabulary. That is, in contrast to previous findings indicating that 19-month-old toddlers fail to generalise word recognition to the non-native accent when words display CS vowel differences [4], even younger 17-month-olds here recognised familiar words produced in an unfamiliar accent that displayed CS vowel differences following pre-exposure to a brief story. Based on these findings, we propose that the pre-exposure passage in either accent provided sufficient contextual support to reduce disruption of cross-accent CS vowel differences in word recognition. This is consistent with other evidence that pre-exposure to phonetic variation in infants' natural speech input impacts on their accommodation of non-native pronunciations, suggesting that variable speech can prepare children to generalise to unfamiliar pronunciations of words [19]. Similarly, our findings showed that exposing infants to a passage in either their native or an unfamiliar accent provides sufficient exposure to reduce the processing cost incurred as a result of unfamiliar accents. The fact that 17-month-olds in the current study outperformed 19-month-olds on a similar task in [4] further supports the hypothesis that sufficient exposure can reduce the negative impact of cross-accent differences.

Bilinguals did not show any significant differences in word recognition abilities compared to monolinguals. This is at odds with [13], which had proposed that bilinguals' reduced exposure to and reduced vocabulary size in each of their languages may lead to delayed cross-accent processing. Possibly, vocabulary size per language affects infants' performance in the more cognitively demanding task of word identification, which was tested in [13], but not in our easier task of word recognition. Compared to word recognition, which just involves recognizing auditory word forms, word identification is reasoned to be a more difficult task since it involves recognising not only the auditory word forms, but also the referent associated with the form. On the other hand, the pre-exposure possibly benefited the bilinguals and brought them up to the level of the monolinguals thereby removing the bilingual disadvantage previously found in [13]. Conversely, our results could also be inconsistent with the possibility that bilingual input confers an *advantage* in cross-accent word recognition. Bilinguals have not yet been tested on these words and this task without story pre-exposure like the 19-month monolinguals in [4]. Perhaps they would have outperformed the monolinguals even without the story pre-exposure, which may have supported the monolinguals sufficiently to form phonological abstractions that assisted in recognising the familiar toddler words in JaME during the test phrase. Future research could examine both groups' performance without pre-exposure.

5. References

[1] Bosch, L., & Sebastian-Galles, N. (2003). Simultaneous bilingualism and perception of a language-specific vowel contrast in the first year of life. *Language and Speech, 46* (2-3), 217-243.

[2] Fennell, C.T., Byers-Heinlein, K., & Werker, J.F. (2007). Using speech sounds to guide word learning: the case of bilingual infants. *Child Development*. *78*, 1510-1525.

[3] Bosch, L., & Sebastian-Galles, N. (1997). Native language recognition abilities in 4-month-old infants from monolingual and bilingual environment. *Cognition, 65,* 33-69.

[4] Best, C. T. & Kitamura, C. (2014). The role of perceptual assimilation in early development of recognition of words spoken in native vs unfamiliar regional accents. *International Conference on Infant Studies*, Berlin, Germany, 3-5 July.

[5] Clarke, C. M., & Garrett, M. (2004). Rapid adaptation to foreign accented English. *J. Acoustical Society America, 116*, 3647–3658.

[6] Best, C. T., Tyler, M. D., Gooding, T. N., Orlando, C. B., & Quann, C. A. (2009). Development of phonological constancy: Toddlers' perception of native- and Jamaican-accented words. *Psychological Science, 20(5)*, 539-542.

[7] Werker, J. F., & Curtin, S. A. (2005). PRIMIR: A developmental framework of infant speech processing. *Language Learning and Development, 1*, 197–234

[8] Mulak, K. E., Best, C. T., Tyler, M. D., & Kitamura, C. (2013). Development of phonological constancy: 19-month-olds, but not 15-month-olds, identify familiar words spoken in a non-native regional accent. *Child Development*, *84*(6), 2064–2078.

[9] Van Heugten, M., & Johnson, E. K. (2014). Learning to contend with accents in infancy: Benefits of brief speaker exposure. *Journal of Experimental Psychology: General, 143*(1), 340–350.

[10] Best, C.T. (2015) "Devil or angel in the details? Perceiving phonetic variation as information about phonological structure" in The phonetics-phonology interface, J. Romero and M. Riera, Eds. Amsterdam, The Netherlands: John Benjamins, 2015, pp. 3-31.

[11] Byers-Heinlein, K., Fennell, C.T., & Werker, J.F. (2013). The development of associative word learning in monolingual and bilingual infants. *Bilingualism: Lang Cognition, 16*(1), 198–205.

[12] Best, C. T., Tyler, M. D., Kitamura, C., & Bundgaard-Nielsen, R. L. (2010). Vocabulary size at 17 months and the emergence of phonological constancy in word recognition across native and nonnative dialects. Presented at the International Conference on Infant Studies, Baltimore, MD.

[13] Mulak, K.E., & Escudero,P. (2016). The development of cross-accent identification of familiar words by monolingual and bilingual infants. Manuscript in preparation.

[14] Best, C. T., Kitamura, C., Pal, M., & Dwyer, A. (2012). Young toddlers recognise non-native Jamaican-accented words differing only in "category-goodness" from native-accent pronunciations. *International Conference on Infant Studies (ICIS)*, Minneapolis MN, USA, June 7-9.

[15] Fennell, C.T., Byers-Heinlein, K., & Werker, J.F. (2007). Using speech sounds to guide word learning: The case of bilingual infants. *Child Development*. *78*, 1510-1525.

[16] Dale, P. S., & Fenson, L. (1996). Lexical development norms for young children. *Behavior Research Methods, Instruments, & Computers, 28*, 125–127.

[17] Kucera, H., & Francis, W.N. (1967). Computational analysis of present day American English. Providence, RI: Brown U Press.

[18] Halle, P.A., & de Boysson-Bardies, B. (1996). The format of representation of recognised words in infants' early receptive lexicon. *Infant Behavior and Development*, *19*, 463–481.

[19] Schmale, R., Cristia, A., & Seidl, A. (2012) Toddlers recognize words in an unfamiliar accent after brief exposure. *Developmental Science, 15* (6), 732-738.

Tailoring language training to prevent cognitive overload and improve phonetic learning outcomes

Dragana Ninkovic, Ammie Hill, and Mark Antoniou

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

18056929@student.westernsydney.edu.au, ammiel.hill@gmail.com,
m.antoniou@westernsydney.edu.au

Abstract

Language training programs often employ one-size-fits-all approaches that do not consider individual differences between learners. Cognitive views of foreign language learning propose that training should be tailored to suit the abilities of individual learners. We compared two approaches to artificial language learning in adults: 1. tailored training that began with passive exposure and adapted based on the learner's performance, versus 2. training with corrective feedback. Tailored training resulted in superior learning to feedback-only, and learning correlated with language aptitude and working memory. We suggest that tailored training programs that take into account individual differences may lead to desirable training outcomes and have pedagogical implications.

Index Terms: phonetic learning, language aptitude, tailoring, individual differences

1. Introduction

Learning a foreign language is cognitively demanding. Phonetic learning is particularly challenging and can affect language acquisition more broadly. It is well known that non-native talkers speak with a foreign accent. Research on cross-language speech perception has also demonstrated that listeners hear with an accent [1]. Work in this area has demonstrated that nonnative speech sounds are perceived in relation to native categories, a process termed assimilation [2]. Theoretical models have been developed to account for these language-specific effects on speech perception [3]. Difficulties in phonetic learning are likely to lead to subsequent difficulties in language acquisition including vocabulary, comprehension, and literacy. Thus, improving phonetic learning will likely benefit language learning more broadly.

Individual learners vary greatly in terms of their cognitive abilities (including memory, intelligence, attention), preferred learning methods, and motivation. However, most language teaching takes a universal approach, assuming (often implicitly) that individuals learn in the same way. This results in a discrepancy between the considerable variation between the needs of learners, and the universal approach taken by most language course books. The aim of the present study was to tailor language training proactively by modifying training based on the learner's level of performance across training sessions.

Successful language learning depends on a complex interaction between training-related and learner-specific factors [4]. Training-related factors refer to characteristics of the learning situation, including the characteristics of the learning task, teaching method, input from native speakers, and whether corrective feedback is provided. Learner-specific factors, on the other hand, refer to characteristics of the individual learner, such as intelligence, motivation, language aptitude, working memory, and prior language experience, including bilingualism. Cognitive approaches to language learning match learners to the correct training type by taking into account such individual differences. One example comes from the speech processing literature, within which it is generally thought that talker variability (i.e., being exposed to the speech of different talkers) results in more robust learning, superior generalisation to novel stimuli, and greater long-term retention of phonetic information [5]. However, questions have been raised concerning the universality of the learning benefit of talker variability. American-English native speakers were asked to learn 18 pseudo-words. Four talkers produced six pseudo-words, each comprised of a single syllable with one of three pitch contours: level, rising or falling. It was found that high talker variability only benefitted learners with strong pre-training abilities. Learners with weak pre-training abilities were actually impaired relative to the single-talker condition [6]. This finding has propelled a line of research examining the interaction between learner-specific and training-related factors in the learning of non-native phonetic contrasts.

The first study of this series [7] compared native English listeners with Mandarin-English and Korean-English bilinguals in their ability to learn three artificial languages that were similar to different natural languages (English-like, Mandarin-like, and Korean-like). The aim was to test if native language knowledge improved learning of a foreign, but familiar, phonetic contrast. The results showed that this was not the case. All learners found the Mandarin-like language easiest to learn, followed by the English-like, and the Korean-like was the most difficult. This suggests that some language features are easier to learn than others, regardless of who is doing the learning.

The next study in this series [8] presented English listeners with an expanded set of five artificial languages (Vietnamese-like and Arabic-like languages were added). The results replicated [7]: Mandarin-like was again the easiest to learn, followed by English-like, however, Korean-like, Vietnamese-like, and Arabic-like were learned equally poorly. Individual differences in language aptitude and working memory correlated with language learning in four of the five artificial languages, suggesting that these variables may separate high from low aptitude learners.

Antoniou and Blair [9] built on this work by examining how individual differences interact with training paradigm design to determine learning of the three equally difficult-to-learn-languages. Training was modified in three ways, one for each difficult-to-learn language: double exposure (passive learning with twice the number of exposure trials), chunking (presenting minimal-pair words in couplets to highlight critical phonetic differences), and corrective feedback (given after each learning trial). High aptitude learners benefited from all training types, whereas low aptitude learners only benefited from double exposure, that is, the least cognitively demanding training type.

Cognitive resource theory offers a possible explanation for these findings. High aptitude learners may have more available cognitive resources to devote to processing corrective feedback,

enabling them to benefit from the additional information provided. Conversely, low aptitude learners may not have spare cognitive resources in reserve and will exhibit a performance decrement when presented with additional information (i.e., chunking or feedback, listed in order of increasing cognitive demands). This explanation is consistent with cognitive approaches to speech processing [10], [11]. It follows that if low aptitude learners have insufficient cognitive resources to dedicate to the processing of additional information, it may be advantageous to delay the introduction of more cognitively demanding training until learners develop some level of competence, aspects of the perceptual process become automatised and sufficient cognitive resources have been freed.

The aim of the present study was to tailor language training proactively over seven sessions, taking into account the needs of individual learners. An adaptive training method was used that increased task demands once learners reached a predetermined performance criterion, and compared the learning outcomes to those from a fixed approach to training (feedback). Specifically, it was hypothesised that:

1. Tailored language training will result in superior learning outcomes than untailored feedback-only training.
2. Language aptitude will positively correlate with performance in the language training program.
3. Working memory will also positively correlate with performance in the language training program.

2. Method

2.1. Participants

Forty-two Australian English native speakers (*M* age = 22.0; *SD* = 4.6) participated in the experiment. Thirty-five were undergraduate psychology students at Western Sydney University. Seven were recruited through advertisements in the community. None reported any history of audiologic or neurological deficits. All passed a pure-tone hearing screening at 25 dB HL at 500, 1000, 2000, and 4000 Hz.

2.2. Materials

Participants were presented with an Arabic-like artificial language that consisted of eight words, which differentiated words using a single, critical phonetic feature: a voiceless velar-uvular ejective contrast /k'/-/q'/. Ejectives are voiceless speech sounds created using a glottalic egressive airstream, where air is compressed by upward movement of the glottis. /k'/ and /q'/ differ according to their places of articulation, which are the soft palate (velum) and the uvula, respectively. The two consonants were produced with four different vowel endings /e/, /i/, /o/, /u/, to generate the eight artificial words that comprised the language. These stimuli were produced by a male native speaker of Quechua, a language that contrasts velar and uvular ejective stops. The recordings were created within a sound attenuated booth, using a Shure SM58 cardioid microphone attached to a boom stand.

2.3. Procedure

Language aptitude was assessed using subtests B, D and E of the LLAMA [12] suite. LLAMA subtest B measures vocabulary learning ability, D measures sound recognition ability, and E measures the ability to form sound-symbol associations. Working memory was assessed using the Verbal Attention test from the Woodcock-Johnson IV Tests of Cognitive Abilities [13]. Participants hear an intermixed series of numbers and animal names, and then must answer a question. For example, participants hear "lamb, 6, pony" and are asked to recall the number between lamb and pony. Sequences increase in difficulty as the test progresses.

The artificial language experiment was presented using Sennheiser HD 280 Pro headphones connected to a HP Pro Book 650 laptop running E-Prime Professional 2 software. Stimulus output level was calibrated to 72 dB SPL.

Participants completed 7 training sessions on separate days with no more than three non-training days between sessions. Two training paradigms were used within this study: passive exposure and corrective feedback, following [8], [9]. During training, each of the eight words were paired with a picture.

Participants were randomly assigned to one of two training groups: 1. Those in the tailored training condition completed passive exposure training in each session until they achieved an accuracy score of 60% in the word identification test before moving to corrective feedback training for the remaining sessions. 2. Control group participants completed corrective feedback training in all sessions.

The passive exposure training paradigm involved 192 trials (8 word-picture pairings × 24 repetitions), with no response required to advance to the next word. Each word-picture pairing was presented randomly at a rate of every 3.5 seconds.

The corrective feedback training paradigm also involved 192 trials. Word-picture pairings were the same as in passive exposure training, but, minimal pairs ending with the same vowel (e.g., /k'i/-/q'i/) were blocked into sets. Each set included 16 trials, comprised of eight exposure trials (four to each word), followed by eight quiz trials during which corrective feedback was given. On quiz trials, each word was auditorily presented and subjects selected the correct picture out of two options. The four sets were repeated three times each, resulting in 192 trials total ([8 exposure + 8 quiz trials] × 4 pairs × 3 repetitions). The order of the sets was randomised.

Following training, participants completed a word identification test that consisted of 64 trials. During test, all eight pictures were presented on-screen. Each word was auditorily presented in random order, and participants selected the corresponding picture via a keypress (1-8). The test was self-paced and no feedback was given. Word identification accuracy scores were calculated upon completion.

3. Results

3.1. Pre-training cognitive assessments

A series of independent samples *t*-tests were conducted to ensure that the tailored and feedback-only groups were matched prior to the commencement of training. The groups did not

Table 1. *Means (and standard deviations) for the tailored and feedback-only groups for the pre-training variables of age, LLAMA language aptitude subtests B, D, E, and Verbal Attention (VA).*

Group	Age	LLAMA B	LLAMA D	LLAMA E	VA
Tailored	22.70 (6.05)	59.75 (25.26)	24.00 (21.50)	84.00 (25.01)	119.30 (4.20)
Feedback-only	21.27 (2.71)	57.27 (23.23)	24.09 (18.75)	84.09 (13.33)	118.55 (3.85)

Table 2. *Pearson's correlations for the tailored and feedback-only groups between language learning across sessions 1-7 and cognitive measures LLAMA language aptitude subtests B, D, E, and Verbal Attention (VA).*

Group	Session	LLAMA B	LLAMA D	LLAMA E	VA
Tailored	1	.68***	.45*	.13	.47*
	2	.73***	.53*	-.09	.61**
	3	.76***	.40	.10	.72***
	4	.66**	.34	.07	.49*
	5	.68**	.35	.20	.54*
	6	.59**	.15	.15	.44
	7	.51*	.26	.15	.43
Feedback-only	1	.67***	.26	.16	.25
	2	.52*	-.02	-.00	.47*
	3	.54**	.10	.08	.54**
	4	.53*	.14	.19	.47*
	5	.53*	.05	.05	.43*
	6	.45*	.02	.17	.51*
	7	.53*	-.03	.21	.46*

differ in age, $t(25.69) = .97$, $p = .34$, or cognitive test scores for LLAMA B, $t(40) = .33$, $p = .74$, LLAMA D, $t(40) = .02$, $p = .99$, LLAMA E, $t(28.37) = .01$, $p = .99$, or Verbal Attention $t(40) = .04$, $p = .97$ (see Table 1). The absence of between-group pre-training differences suggests that any observed differences in artificial language learning are likely due to experimental manipulations.

3.2. Language learning

Accuracy scores were compared via a 2 × (7) ANOVA, with the between-subjects factor of training group (tailored vs. feedback-only) and the within-subjects factor of session (1-7).

There was a significant main effect of training group, $F(1,40) = 7.17$, $p = .01$, $\eta_p^2 = .15$, indicating that the participants who received tailored training ($M = 66.1$, $SD = 19.0$) outperformed those who received feedback-only training ($M = 51.3$, $SD = 17.6$). There was also a significant main effect of session with a Greenhouse-Geisser adjustment to the degrees of freedom $F(2.34, 93.45) = 35.52$, $p < .001$, $\eta_p^2 = .47$, indicating that performance improved across training sessions, overall.

Importantly, there was a significant Training Group × Session interaction with a Greenhouse-Geisser adjustment to the degrees of freedom, $F(2.34, 93.45) = 7.21$, $p < .001$, $\eta_p^2 = .15$. This suggests that the two groups showed different patterns of learning across the seven sessions. As shown in Figure 1, while the two groups performed similarly initially, they diverged as they progressed through training sessions.

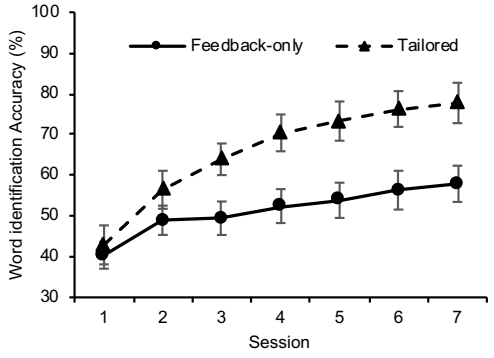


Figure 1: *Mean word identification accuracy for feedback-only and tailored training groups across seven training sessions. Error bars represent standard error of the mean.*

This was further explored in a series of post-hoc independent-samples *t*-tests, one for each of the seven training sessions. A Bonferroni-adjusted α level of .007 (.05 / 7) was employed to control for multiple comparisons. The training groups did not significantly differ in their learning performance in session 1, $t(40) = .42$, $p = .69$, $d = 0.13$, or session 2 $t(40) = .52$, $p = .60$, $d = 0.16$. By session 3, the training groups began to show signs of divergence in their learning performance, but this was not statistically significant at the Bonferroni-adjusted α level of .007, $t(40) = 2.57$, $p = .01$, $d = 0.79$. From session 4 onwards, the tailored training group exhibited superior learning performance relative to the feedback-only group, as evidenced by significant *t*-tests for session 4, $t(40) = 3.06$, $p = .004$, $d = 1.35$, session 5, $t(40)$ 3.19, $p = .003$, $d = 0.98$, session 6, $t(40) = 3.19$, $p = .003$, $d = 0.98$, and session 7, $t(40) = 3.23$, $p = .002$, $d = 0.99$.

3.3. The roles of language aptitude and working memory

Bivariate correlations were conducted to explore the association between the cognitive measures (language aptitude and working memory) and language learning (see Table 2). LLAMA B positively correlated with language learning across all seven sessions for both the tailored and feedback-only training groups. This suggests that LLAMA B may be a useful measure for identifying high and low language aptitude groups for language learning programs. Similarly, verbal attention scores positively correlated with learning in most sessions for both groups, with the exception of sessions 6 and 7 for the tailored group, and session 1 for the feedback-only group. LLAMA subtests D and E did not correlate as robustly with learning across the seven sessions for either group.

4. Discussion

The present study examined the effectiveness of an individually tailored training program as compared to a one-size-fits-all approach to language learning. This was done by implementing a multisession training paradigm which adaptively increased cognitive demands according to an individual's performance (tailored training group). Tailored training began with passive-exposure and switched to corrective feedback training only once the 60% word learning criterion was met. The feedback-only control group completed the same training with corrective feedback for all sessions regardless of performance. As predicted, tailored training resulted in superior language

learning outcomes. The two training groups performed comparably across the first two sessions, but began to diverge by session three, with the tailored group exhibiting better performance than the feedback-only from session 4 onwards.

The findings are consistent with prior work that has emphasised the importance of matching speech training paradigms to the cognitive abilities of learners [6], [9]. High aptitude learners are able to benefit from more sophisticated (but also cognitively demanding) training paradigms, whereas low aptitude learners benefit from training that keeps cognitive demands low, such as passive exposure, and may actually be hindered by cognitively demanding training methods, such as corrective feedback [9]. Similarly, we have observed that once learners achieve a level of mastery, they may benefit from more sophisticated (and cognitively demanding) training. This pattern of findings is accounted for by the capacity theory of comprehension [11], according to which learners have a finite pool of cognitive resources that they may dedicate to specific language learning tasks. During the initial stages of learning, it may be desirable to keep additional cognitive demands low. With exposure, and as perceptual processes become automatised, cognitive resources are freed and these may be dedicated to more sophisticated language learning methods.

For both groups, language learning correlated with language aptitude (vocabulary learning ability as measured by LLAMA B), as well as working memory. Working memory correlated less robustly with language learning than did language aptitude, suggesting that language aptitude may be a more useful predictor of language learning. During the initial stages of learning, correlations appeared to be more robust in the tailored training paradigm. This lends support to the capacity theory interpretation described above. The more demanding feedback-only training method may have initially overwhelmed some learners and had a detrimental effect on learning as evidenced by the flatter curve in Figure 1.

The implications of these results are that one-size-fits-all language learning programs may not be beneficial for all learners, and thus not as effective as adaptive programs that can be tailored to the needs of individuals [6]. We would predict improvements in learning performance when cognitive demands are reduced during the initial stages of learning, which should be especially beneficial for poor learners. By initially implementing a low demand task, participants are familiarised with the novel language without placing additional cognitive demands upon them. During this passive exposure phase, processing becomes automatised and cognitive resources are freed. It is posited that by freeing cognitive resources in this manner, learners are then able to benefit from the additional information presented in the more cognitively demanding corrective feedback training type. The findings in the current study indicate that tailoring training based on cognitive demand is a successful strategy to improve overall learning outcomes.

These findings are likely to inform language training paradigms. Although informative, the findings also raise many new questions concerning how language training can be optimised. One limitation of this study is that we cannot rule out the possibility that passive-exposure-only training would have resulted in superior learning outcomes to feedback-only training because the present study did not include a passive-exposure-only control group. This possibility seems unlikely because training with corrective feedback is generally considered to result in the best training outcomes. Nevertheless, it would be beneficial to compare tailored training to passive-exposure-only and feedback-only control groups. Additionally, it would be informative to investigate how multiple training paradigms of differing cognitive demands can be combined to determine optimal language training approaches. It would also be informative to extend our tailored approach to the learning of other nonnative phonetic contrasts and linguistic features.

5. Conclusions

In conclusion, this study has found that an individually tailored training program results in greater language learning than a one-size-fits-all approach to language learning. These results demonstrate that individual differences in the availability of cognitive resources interact with task variables to determine learning success. Overall, these findings will inform the design and implementation of tailored approaches to language training based on individual cognitive profiles, and training methods.

6. Acknowledgements

This study was funded by Australian Research Council Discovery Early Career Research Award DE150101053 to MA.

7. References

[1] Antoniou, M., Tyler, M. D. and Best, C. T., “Two ways to listen: Do L2-dominant bilinguals perceive stop voicing according to language mode?” J. Phon., 40:582–594, 2012.

[2] Antoniou, M., Best, C. T. and Tyler, M. D., “Focusing the lens of language experience: Perception of Ma’di stops by Greek and English bilinguals and monolinguals,” J. Acoust. Soc. Am., 133:2397–2411, 2013.

[3] Best, C. T. and Tyler, M. D., “Nonnative and second-language speech perception: Commonalities and complementarities,” in O.-S. Bohn and M. J. Munro, [Eds.], Language experience in second language speech learning: In honor of James Emil Flege, 17:13–34, John Benjamins, 2007.

[4] Wong, P. C. M. and Ettlinger, M., “Predictors of spoken language learning,” J. Commun. Disord., 44:564–567, 2011.

[5] Logan, J. S., Lively, S. E. and Pisoni, D. B., “Training Japanese listeners to identify English /r/ and /l/: A first report,” J. Acoust. Soc. Am., 89:874–886, 1991.

[6] Perrachione, T. K., Lee, J., Ha, L. Y. Y. and Wong, P. C. M., “Learning a novel phonological contrast depends on interactions between individual differences and training paradigm design,” J. Acoust. Soc. Am., 130:461–472, 2011.

[7] Antoniou, M., Liang, E., Ettlinger, M. and Wong, P. C. M., “The bilingual advantage in phonetic learning,” Biling. Lang. Cogn., 18:683–695, 2015.

[8] Antoniou, M. and Low, T., “Individual differences in language training,” Presented at CoEDL, Western Sydney University, 2016.

[9] Antoniou, M. and Blair, M., “Tailoring phonetic learning to the needs of individuals on the basis of language aptitude,” in Proceedings of the Sixteenth Australasian International Conference on Speech Science and Technology, 37–40, 2016.

[10] Kahneman, D., Attention and effort, Prentice-Hall, 1973.

[11] Just, M. A. and Carpenter, P. A., “A capacity theory of comprehension: Individual differences in working memory,” Psychol Rev, 99:122–149, 1992.

[12] Meara, P., “LLAMA language aptitude tests: The manual,” 2005.

[13] Schrank, F. A., Mather, N. and McGrew, K. S., “Woodcock-Johnson IV Tests of Cognitive Abilities,” Riverside, 2014.

Tone training for native speakers of tonal and nontonal languages

Jessica L. L. Chin, Mark Antoniou

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

j.chin/m.antoniou@westernsydney.edu.au

Abstract

Nonnative linguistic features can be difficult for adults to learn. Past studies have shown that learners from various language backgrounds can be trained to perceive lexical tone. Here, we trained native speakers of tone languages Mandarin Chinese and Vietnamese and nontonal English to learn the tones of Hakka Chinese. Although tone language native speakers performed better in the first session, all groups improved in performance by session 5. Findings indicate that tone language experience facilitates the learning of new tone categories, but native tonal and intonational categories also contribute to how well specific nonnative tones are learned.

Index Terms: Hakka Chinese, tone training, language background

1. Introduction

As we age, it becomes increasingly difficult to acquire nonnative linguistic features. The native language influences how nonnative speech sounds are perceived such that some nonnative distinctions are difficult to discern. With training, perception of nonnative linguistic features can improve.

Lexical tone is a linguistic feature that distinguishes the meanings of words using pitch. Over half of the world's languages are tone languages [1]. Tone languages vary in the number of tones within their system, and tones may vary in their height, direction and pitch trajectory. Level tones retain a consistent height and direction, while contour tones change in pitch direction along the syllable [2]. Checked tones are shorter in duration and are present in tone languages where stop consonant codas are permissible [3], [4], such as Hakka Chinese.

Studies have shown that tone language experience modulates perception of nonnative tone contrasts [5], [6]. For tone language speakers, native tone categories may facilitate but also interfere with the perception of nonnative tones [7]–[9]. To account for this, theories of cross-language speech perception such as the Speech Learning Model (SLM) [10] and Perceptual Assimilation Model (PAM) [11] have been applied to tone perception, but more research is required to account for assimilation patterns of native listeners of both tonal and nontonal languages [8], [12], [13]. It is not well understood how the *complexity* of the native tone system influences nonnative tone perception. While complexity in the past has been defined as the number of tones within a tone system [14], similarity of pitch slopes and the presence of level, contour and/or checked tones might also contribute to a tone language's complexity. For nontonal language speakers, intonational patterns can facilitate perception of nonnative tones [15], [16].

Tone training studies have observed improvements in tone learning in naive listeners, as well as learners of a tonal L2 with either a tonal or nontonal L1 [17]. There is evidence that native tonal language speakers possess an advantage over nontonal language speakers when learning nonnative tones, but improvements have been found in both groups [5]. However, learners with a tonal or pitch-accented L1 do not always show an advantage [9]. Whether a learner's native language is tonal or nontonal influences their sensitivity to either early or late pitch differences within the syllable [18], and their bias towards pitch height or direction [19]. Further, ease of tone learning is determined by how nonnative tones map onto L1 tonal or intonational categories [15]. Keeping cognitive demands low during training benefits (especially poor) learners [20], [21].

The aims of the present study were to assess if tone language experience facilitates nonnative tone learning, and whether experience with a more complex tone system provides an additional benefit. We compared native speakers of nontonal (Australian English) and tonal languages (Mandarin Chinese and Vietnamese). Learners were presented with tones from Meixian Hakka, a Chinese dialect spoken in southern China, Hong Kong, Taiwan, and South-East Asia [3], [4]. Hakka has four regular tones and two checked tones [4]. Tone 1 (33) is a mid-level tone, tone 2 (11) is low-level, tone 3 (41) is mid-falling, and tone 4 (51) is high-falling [3]. The first checked tone (55) is high-level, and the second (41) is mid-falling. Permissible stop codas in Hakka include /p/, /t/ and /k/, and can appear in the VC or CVC syllable context [3].

Mandarin has four tones: high-level (55), mid-rising (35), low-dipping (214) and high-falling (51) [22], and no checked tones. Southern Vietnamese has five tones: mid-level, low-falling, mid-rising, low falling-rising and falling-rising. Checked versions of the mid-rising and low falling-rising tones appear when a syllable ends in either /p/, /t/ or /k/ [23].

It was hypothesised that both tonal learner groups would outperform the Australian English group, for whom lexical tone is a nonnative feature. Although native intonational patterns can be mapped to nonnative tonal contrasts [15], [24], Hakka's two falling tones and its lack of a rising tone may cause difficulties for English speakers who utilise rising–falling intonational patterns in questions and statements. If tone system complexity predicts learning, the Vietnamese listeners should outperform the Mandarin. However, if similarity to native tone categories predicts learning, a more complex pattern should emerge: Vietnamese speakers should distinguish Hakka's falling tone 3–4 contrast more successfully since a similar contrast occurs in their language [23].

2. Method

2.1. Participants

Participants were 13 native Australian English (AusE) speakers (*M* age = 21.2, SD = 4.6), 16 native Mandarin speakers (*M* age = 26.6, SD = 5.4), and 10 native Vietnamese speakers (*M* age = 27.3, SD = 9.0). None reported any audiologic or neurological deficits. All passed an air conduction audiogram at 25 dB HL at 500, 1000, 2000 and 4000 Hz. Three AusE participants spoke

another language at home, but not a tone language. The Vietnamese group all spoke the Southern dialect, except for three Central dialect speakers. Six Mandarin participants were proficient in another Chinese dialect, but none reported past experience with Hakka. These dialects included Cantonese, Shanghainese, Wuxi, Minnan, and Sichuanese.

2.2. Stimuli

For the tone identification task, a female AusE talker produced the monophthongs /a/, /e/, /i/, /o/, /u/ and the VC syllables /ak/, /ip/, /et/. The syllable context for the checked tone stimuli was kept simple as the focus of the task was to identify the pitch contour of a stimulus. The training and generalisation task stimuli were produced by different talkers and consisted of the CVC nonwords [fon], [leng], [nun], [wa$\bar{p}$], and [mi$\bar{p}$], with the latter two used for the checked tones. All stimuli were nonsense words in English, Mandarin, and Vietnamese, and all segments occurred in the phonemic inventories of each language.

Stimuli were produced with a level tone and were recorded in a sound-attenuated booth using a Shure SM10A cardioid microphone and a Roland Duo-Capture EX audio interface. The stimuli were sampled at 44.1 kHz.

To ensure that the stimuli only differed in pitch, Praat's pitch-synchronous overlap and add function was used to superimpose tone values from three female Hakka speakers [3] onto the recordings for each speaker. Two native speakers of Hakka judged the final stimuli as acceptable Hakka tones.

2.3. Procedure

Participants completed five sessions of training, held on separate days. In session 1, participants completed a demographic questionnaire, a tone identification pre-test, and their first training session. Participants repeated the training task in sessions 2 to 4. In session 5, they completed their final training session, a generalisation task and a tone identification post-test. The experimental tasks were presented on a laptop running E-Prime 3.0, and audio stimuli were presented at 72 dB SPL using Sennheiser HD280 Pro headphones.

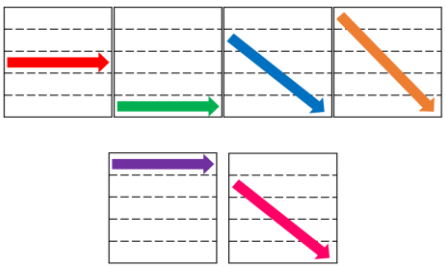


Figure 1: *Arrows visualising unchecked (top) and checked (bottom) Hakka tones used in the tone identification task.*

2.3.1. *Tone identification*

The tone identification task was split into familiarisation, practise and test phases. During familiarisation, participants heard 3 repetitions of /a/ in each of the four tones and /ak/ in the two checked tones. Each sound was accompanied by an arrow visualising the height and direction of the tone heard (see Figure 1). Participants were instructed to memorise this picture-sound association. During practise, participants heard an auditory stimulus and selected the arrow that represented the tone they heard. Feedback was given after each trial. The test phase was similar to the practise, except that no feedback was provided and participants heard previously unencountered stimuli. Additionally, participants were required to respond within 3 seconds. Each sound was presented with every possible response combination, for a total of 104 test trials. The checked and unchecked tone stimuli in all three phases were presented at random in separate blocks. Performance in the tone identification tasks was calculated and expressed as mean percent accuracy. Confusion matrices were created from the proportion of responses made for each tone.

2.3.2. *Training*

Training involved exposure and test. During exposure, participants heard 4 repetitions of a minimal pair (e.g., [nun1] and [nun2]), and saw the image associated with each sound. Participants were then quizzed and asked to choose the correct image from two options for each heard sound. Feedback was provided for each response. During test, participants heard a sound and chose the corresponding picture from all 16 words. In total, there were 96 test trials (16 words × 4 repetitions).

2.3.3. *Generalisation*

Participants completed the generalisation test on their fifth session immediately after completing the training task. This task assessed participants' ability to adapt to speech from a novel talker. The generalisation test was otherwise identical to the test phase of the training task, with a total of 96 trials.

3. Results

3.1. Tone identification pre-test vs. post-test performance

Tone identification scores were compared via a 3 × (2) ANOVA with language as the between-subjects factor and test as the within-subjects factor. As shown in Figure 2, there was a main effect of language on tone identification, $F(2, 36) = 10.12$, $p < .001$, $\eta_p^2 = .360$. Sidak post-hoc comparisons revealed that Mandarin speakers significantly outperformed AusE speakers ($M_{Diff} = 25.39$, 95% CI [11.25, 39.51], $p < .001$). Vietnamese speakers only showed marginally higher performance than AusE speakers ($M_{Diff} = 14.76$, 95% CI [-1.17, 30.68], $p = .076$), and there were no significant differences between Mandarin and Vietnamese speakers ($M_{Diff} = 10.64$, 95% CI [-4.63, 25.90], $p = .245$). There was no main effect of test, $F(1, 36) = 2.415$, $p = .129$, $\eta_p^2 = 0.63$, nor was there an interaction between test and language $F(2,36) = 0.010$, $p = .990$, $\eta_p^2 = .063$.

Confusion matrices were created to examine the spread of responses made by learners in the tone identification tasks (Table 1). All learner groups consistently identified the correct tone, but tonal language speakers were the most accurate. Additionally, the groups' confusion patterns differed. AusE speakers identified level tones 2 and 5 most successfully; they struggled with the other tones (61-64% accuracy). Mandarin speakers identified falling tones 3, 4 and 6 slightly less accurately (81-87%), while Vietnamese speakers misidentified tones 2, 3 and 6 more than other tones (72-74%). Accuracy decreased further for certain contrasts. For instance, AusE speakers identified tones 1 and 3 only 56% and 48% of the time when distinguishing between the two, and tones 3 and 4 64% and 52% of the time. Mandarin speakers correctly identified tones 3 and 4 70% and 69% of the time. When Vietnamese speakers were presented with the tone 2–3 contrast, accuracy in tone 3 did not change (73%), but tone 2 accuracy dropped to 63%. They identified tones 3 and 4 73% and 76% of the time, similar to their overall identification accuracy for both tones.

Post-test	Correct tone																	
Response (%)	T1 (33)			T2 (11)			T3 (41)			T4 (51)			T5 (55)			T6 (41)		
	E	M	V	E	M	V	E	M	V	E	M	V	E	M	V	E	M	V
T1 (33)	**64**	**94**	**82**	7	2	7	16	3	11	15	2	9						
T2 (11)	10	3	5	**80**	**91**	**73**	12	7	8	9	3	5						
T3 (41)	13	1	7	5	5	13	**63**	**81**	**72**	15	9	7						
T4 (51)	14	2	6	8	2	8	9	9	9	**61**	**86**	**78**						
T5 (55)													**70**	**94**	**79**	37	13	26
T6 (41)													30	6	21	**63**	**87**	**74**

Table 1. *Tone confusion matrix for English (E), Mandarin (M), and Vietnamese (V) groups in the tone identification post-test. Correct responses are indicated in boldface.*

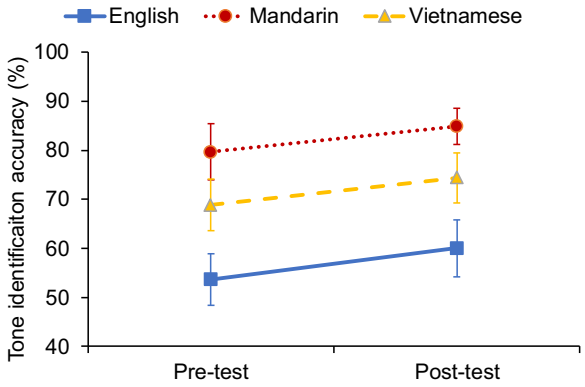


Figure 2: *Tone identification accuracy (%) at pre- and post-test. Error bars depict SEM.*

3.2. Tone training and generalisation performance

Tone training scores were submitted to a 3 × (5) ANOVA with language as the between-subjects factor and session as the within-subjects factor. There was a main effect of language, $F(2, 36) = 12.89$, $p < .001$, $\eta_p^2 = .417$. Sidak post-hoc comparisons revealed that the Mandarin ($M_{Diff} = 30.91$, 95% CI [15.60, 46.23], $p < .001$) and Vietnamese groups ($M_{Diff} = 20.02$, 95% CI [2.77, 37.27], $p = .019$) both outperformed the AusE group, and the Mandarin and Vietnamese groups did not differ ($M_{Diff} = 10.90$, 95% CI [-5.64, 27.43], $p = .289$). There was also a main effect of session, $F(2.56, 92.09) = 86.82$, $p < .001$, $\eta_p^2 = .707$. Performance in sessions 1 to 4 differed significantly ($p < .001$; sessions 1 and 3, $p = .008$), but there was no difference between session 4 and 5 (p = .07). There was no interaction between session and language $F(5.12, 86.82) = 1.67$, $p = .148$, $\eta_p^2 = .085$.

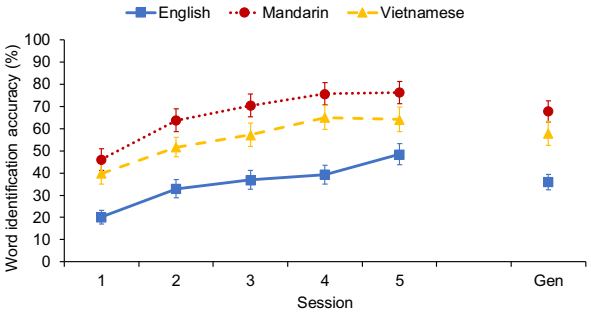


Figure 3: *Tonal (Mandarin, Vietnamese) and nontonal (AusE) learners' word identification accuracy (%) across five training sessions and generalisation test (Gen).*

Generalisation test scores were submitted to a one-way ANOVA with language as the between-subjects factor. A main effect of language was observed, $F(2, 36) = 13.64$, $p < .001$, $\eta_p^2 = .431$. Sidak post-hoc analyses found that both the Mandarin ($M_{Diff} = 31.80$, 95% CI [16.40, 47.19], $p < .001$) and Vietnamese ($M_{Diff} = 21.94$, 95% CI [4.60, 39.28], $p = .009$) groups significantly outperformed the AusE group. The Mandarin and Vietnamese groups did not differ ($M_{Diff} = 9.86$, 95% CI [-6.76, 26.48], $p = .378$). Mean scores for all training sessions and the generalisation test are shown in Figure 3.

4. Discussion

In this study, tonal (Mandarin, Vietnamese) and nontonal (AusE) listeners completed five sessions of Hakka tone training. Tone identification ability was measured at pre- and post-test, and tone word learning ability was monitored across the five training sessions. Across all tasks in session 1, the tonal language groups outperformed the nontonal AusE group, showing an initial performance boost for tone language native speakers. This likely reflects the crucial role of tone in tone listeners' perceptual development [25]. Additionally, the tonal groups showed greater ability in generalising to a novel talker than the AusE group. However, the rate of improvement across groups was similar. Only a marginally significant difference was discovered between Vietnamese and AusE listeners in tone identification. Note that data collection is ongoing, and this difference may become statistically significant as power increases. No significant difference was observed between pre- and post-test performance, which suggests that while listeners were able to learn tonal contrasts and associate nonnative sounds to images, training did not improve listeners' ability to identify the tones' pitch slopes.

The two tonal groups showed comparable learning of nonnative Hakka tones, suggesting that the native tone learning advantage is not modulated by the complexity of the native tone system. This finding does not support claims that tonal complexity leads to greater novel tone learning [14], [26].

Nontonal language speakers may map novel tones to native question and statement intonational categories [24], but tones that are not perceived as sufficiently similar to native intonational categories are often identified poorly [15]. In this study, the lack of a rising tone and the presence of two falling tones may have hindered the AusE group's learning performance, as the Hakka tones were too dissimilar to their native rising and falling intonation patterns. Interestingly, AusE learners identified tones 2 (low-level) and 5 (high-level) most accurately. This may be attributed to greater weighting toward pitch height in English. One possible explanation is that AusE listeners found tone 5 (a checked tone) easy to identify because the only other available response option was tone 6, but the AusE group performed as poorly in identifying tone 6 as they

did tone 4 (both falling 41 tones). It has been posited that English listeners are initially able to distinguish only two tones by pitch height, but the AusE group's lower identification accuracy for the mid-level tone 1 in the post-test is not consistent with the high identification accuracy observed for a similar Cantonese tone [15]. It appears that the AusE learners in this study were only able to distinguish between high- and low-level tone contrasts.

Regardless, all language groups improved significantly during training. Though AusE speakers did not perform as well as the tonal groups, their significant improvement in performance supports the notion that novel linguistic features can be learned in adulthood through speech training protocols [5], [6], [9], [15].

Further examination of confusion patterns revealed language-specific differences in tone identification which can be addressed by the principles in PAM [11]. Although overall tone identification accuracy in both pre- and post-test was high across all groups, some contrasts were more difficult to distinguish than others. As predicted, Mandarin speakers showed slightly poorer identification accuracy when exposed to the falling tone 3–4 contrast since they only have one native falling tone category, but their accuracy was high enough to indicate they may have been able to categorise both tones as different tones. English speakers were not able to distinguish the 1–3 and 3–4 contrasts, judging from their low identification accuracy of the contrasts. These particular tones do not map closely enough onto any native intonational categories, leading to poorer identification [15]. Tone 3 appears to have been assimilated into the tone 1 category as a poor exemplar of tone 1. A similar pattern was seen in Vietnamese learners' identification of the tone 2–3 contrast, where only tone 2 accuracy decreased. The Vietnamese tone system does not have a low-level tone, but it does have two falling tones [23], which explains the poorer identification of tone 2 in the 2–3 contrast but unchanged identification accuracy of the tone 3–4 contrast. Future training studies could implement discrimination and category-goodness rating tasks to determine the assimilation patterns for these contrasts.

5. Conclusions

The current study assessed the effect of language background on the learning of a new tone system. While native speakers of nontonal and tonal languages were able to benefit from tone training, tonal language speakers showed an advantage in the first session. Although higher performance by the Vietnamese group compared to the AusE group was only marginally significant in the tone identification task, the difference between both groups may reach statistical significance with further data collection. Preliminary findings already suggest that the complexity of the native tonal language does not provide additional advantages for nonnative tone learning. The learning of nonnative tones seems to depend on how well they map to an individual's native tonal or intonational categories. Further investigation on the assimilation patterns for different language groups is required. These findings have implications for theories of nonnative phonetic learning.

6. References

[1] Yip, M., Tone, Cambridge University Press, 2002.

[2] Abramson, A. S., "Static and dynamic acoustic cues in distinctive tones," Lang. Speech, 21:319–325, 1978.

[3] Cheung, Y. M., "Vowels and tones in Meixian Hakka: An acoustic and perceptual study," City University of Hong Kong, 2011.

[4] Hashimoto, M. J., The Hakka dialect: A linguistic study of its phonology, syntax and lexicon, Cambridge University Press, 2010.

[5] Wayland, R. P. and Li, B., "Effects of two training procedures in cross-language perception of tones," J. Phon., 36:250–267, 2008.

[6] Wayland, R. and Guion, S., "Perceptual discrimination of Thai tones by naive and experienced learners of Thai," Appl. Psycholinguist., 24:113–129, 2003.

[7] Bent, T., Bradlow, A. R. and Wright, B. A., "The influence of linguistic experience on the cognitive processing of pitch in speech and nonspeech sounds," J. Exp. Psychol. Hum. Percept. Perform., 32:97–103, 2006.

[8] So, C. K. and Best, C. T., "Cross-language perception of non-native tonal contrasts: Effects of native phonological and phonetic influences," Lang. Speech, 53:273–293, 2010.

[9] Wang, X., "Perception of Mandarin tones: The effect of L1 background and training," Mod. Lang. J., 97:144–160, 2013.

[10] Flege, J. E., "Second language speech learning: Theory, findings and problems," in W. Strange, [Ed.], Speech perception and linguistic experience: Issues in cross-language research, 233–277, York Press, 1995.

[11] Best, C. T., "A direct realist view of cross-language speech perception," in W. Strange, [Ed.], Speech perception and linguistic experience: Issues in cross-language research, 171–204, York Press, 1995.

[12] Wu, X., Munro, M. J. and Wang, Y., "Tone assimilation by Mandarin and Thai listeners with and without L2 experience," J. Phon., 46:86–100, 2014.

[13] Reid, A. et al., "Perceptual assimilation of lexical tone: The roles of language experience and visual information," Atten. Percept. Psychophys., 77:571–591, 2015.

[14] Tong, X. and Tang, Y. C., "Modulation of musical experience and prosodic complexity on lexical pitch learning," presented at Speech Prosody, Boston, USA, 217–221, 2016.

[15] Francis, A. L., Ciocca, V., Ma, L. and Fenn, K., "Perceptual learning of Cantonese lexical tones by tone and non-tone language speakers," J. Phon., 36:268–294, 2008.

[16] Hallé, P. A., Chang, Y.-C. and Best, C. T., "Identification and discrimination of Mandarin Chinese tones by Mandarin Chinese vs. French listeners," J. Phon., 32:395–421, 2004.

[17] Antoniou, M. and Chin, J. L. L., "What can lexical tone training studies in adults tell us about tone processing in children?," Front. Psychol., 9:1–12, 2018.

[18] Kaan, E., Wayland, R., Bao, M. and Barkley, C. M., "Effects of native language and training on lexical tone perception: An event-related potential study," Brain Res., 1148:113–122, 2007.

[19] Chandrasekaran, B., Sampath, P. D. and Wong, P. C. M., "Individual variability in cue-weighting and lexical tone learning," J. Acoust. Soc. Am., 128:456–465, 2010.

[20] Antoniou, M. and Wong, P. C. M., "Poor phonetic perceivers are affected by cognitive load when resolving talker variability," J. Acoust. Soc. Am., 138:571–574, 2015.

[21] Perrachione, T. K., Lee, J., Ha, L. Y. Y. and Wong, P. C. M., "Learning a novel phonological contrast depends on interactions between individual differences and training paradigm design," J. Acoust. Soc. Am., 130:461–472, 2011.

[22] Chen, M. Y., Tone Sandhi: Patterns across Chinese dialects, Cambridge University Press, 2000.

[23] Brunelle, M., "Tone perception in Northern and Southern Vietnamese," J. Phon., 37:79–96, 2009.

[24] Braun, B. and Johnson, E. K., "Question or tone 2? How language experience and linguistic function guide pitch processing," J. Phon., 39:585–594, 2011.

[25] Antoniou, M., To, C. K. S. and Wong, P. C. M., "Auditory cues that drive language development are language specific: Evidence from Cantonese," Appl. Psycholinguist., 36:1493–1507, 2015.

[26] Zheng, H.-Y., Minett, J. W., Peng, G. and Wang, W. S.-Y., "The impact of tone systems on the categorical perception of lexical tones: An event-related potentials study," Lang. Cogn. Process., 27:184–209, 2012.

Factors affecting talker adaptation in a second language

Anne Cutler, L. Ann Burchfield, and Mark Antoniou

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

a.cutler/a.burchfield/m.antoniou@westernsydney.edu.au

Abstract

Listeners adapt rapidly to previously unheard talkers by adjusting phoneme categories using lexical knowledge, in a process termed lexically-guided perceptual learning. Although this is firmly established for listening in the native language (L1), perceptual flexibility in second languages (L2) is as yet less well understood. We report two experiments examining L1 and L2 perceptual learning, the first in Mandarin-English late bilinguals, the second in Australian learners of Mandarin. Both studies showed stronger learning in L1; in L2, however, learning appeared for the English-L1 group but not for the Mandarin-L1 group. Phonological mapping differences from the L1 to the L2 are suggested as the reason for this result.

Index Terms: speech perception, perceptual learning, Mandarin, English, second language learning

1. Introduction

Human listeners adapt to newly-encountered talkers with remarkable rapidity. In the past decade and a half, this process has been extensively investigated using a paradigm in which listeners hear ambiguous phonetic forms which they are able to disambiguate by reference to existing knowledge (see [1] for a review). The initial use of this paradigm [2] established that exposure to just 20 instances of a deviant phonemic form induces learning about the speaker's putative pronunciation of the sound in question, as long as the deviant form is heard in real-word contexts so that it can be ascribed to a phonemic category. Thus an ambiguous sound between /s/ and /f/ will be learned as /s/ if heard in words like *horse*, as /f/ if heard in words like *giraffe*, but will remain ambiguous if heard in nonwords such as *liff* or *liss*. The learning generalises to other words containing the same phoneme, creating a path for rapid adaptation on first exposure to speech from a new talker.

The perceptual learning process has been documented for numerous types of phoneme, in different positions in the word, and for listeners from childhood to old age. It has been shown to be rapid (less than 20 exemplars also work), in good part speaker-specific though with reasonable generalisations across phoneme class and language variety, and long-lasting; further, it can be induced by exposure to words in context or in isolation, and can be measured in either phonetic decision tasks or lexical disambiguation tasks (on all counts see [1]).

Of particular relevance to the present work is that this rapid learning has been observed in many languages, both European and non-European (and indeed that it also holds for lexically distinctive non-segmental speech sounds; in Mandarin, for example, a speaker's use of a lexical tone ambiguous between Mandarin tones 1 and 2 led to adjustment [3] in the same way as was also seen for ambiguous phonemes in Mandarin [4]). Furthermore, the learning can be successfully applied in a second language [5,6], although this has chiefly been observed in related L1 and L2 and in an immersion environment.

In the present study we addressed the relative ability of listeners to apply this rapid adjustment in their first and second languages, and we chose a pair of languages with a high degree of phonological dissimilarity: Mandarin Chinese and (Australian) English. Perceptual learning has already been demonstrated for these languages [4,7], so we expect each L1 to exhibit a robust effect. However, the L2 cases may differ. One of these is an immersion environment: Experiment 1, in which we examined the learning effect in the L2 of Mandarin-English late bilinguals living in Australia. The other is a case without immersion: Experiment 2, where we examined the same learning for L2 in Australian adult learners of Mandarin, still resident in their English-speaking native environment.

2. Pilot experiment

To select an ambiguous fricative for the main experiments, a pilot experiment was conducted in which participants heard and categorised steps from an [f]-[s] continuum in English and Mandarin. A female native speaker of each of Mandarin and English produced the syllables /fu/, /su/, and /θu/. The fricative portions of the /fu/ and /su/ recordings were excised and a 41-step continuum was created in each language (following [2]). Using Praat [8], [f] and [s] waveforms were mixed in constant proportions along a 41-step continuum such that one end of the continuum was 100% [f], 0% [s] and the other end 0% [f], 100% [s]. Fricatives were spliced onto the vowel /u/ taken from the same speaker's production of /θu/. This avoided coarticulatory cues in vowels biasing listeners to interpret the ambiguous sounds as either [f] or [s]. Fourteen steps were chosen from this [f]-[s] continuum as stimuli for this pilot: 1 ([f]), 7, 9, 11, 13, 15, 17, 19, 21, 23, 25, 27, 29 and 41 ([s]). For each continuum (Mandarin, English), a separate group of native listeners heard these 14 steps 10 times in random order and categorised each token by pressing "F" or "S" on a computer keyboard.

Mandarin and English results from the pilot are shown separately in Figure 1. In both languages, listeners' responses proved step 17 to be the most ambiguous token of the 14 steps tested, and thus step 17 was in each case used to construct the ambiguous stimuli for the perceptual learning training trials.

3. Experiment 1

3.1. Method

3.1.1. Participants

24 Mandarin-English late bilingual speakers (mean age 26.6, range = 21.6-36.3; mean age of arrival in Australia 23.3, range = 15.2-35.4; mean length of residence 3.2 years, range = 0.5-9.6) took part in return for a small payment. All reported Mandarin Chinese as their dominant language, and did not report pre-school-age exposure to other languages or dialects. None reported any vision, hearing, or language impairments.

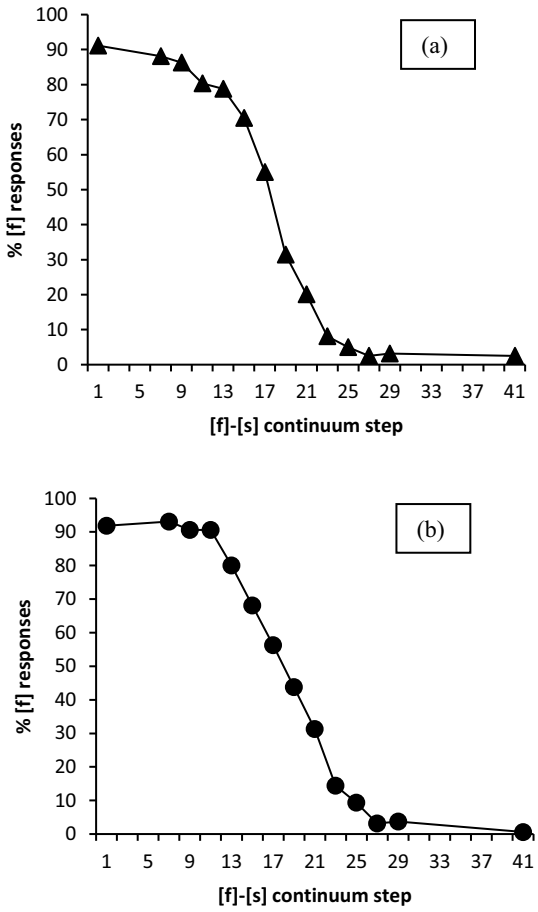


Figure 1: *Pilot experiment: Total proportion of [f] responses for (a) Mandarin and (b) English. In each case, step 17 was chosen as the ambiguous sound for training trials.*

3.1.2. Training stimuli

The training materials for each language were all disyllabic: 100 words and 100 non-words. 60 words were filler items and 40 were training items. Half of the latter were f-words (with [f] as first phoneme of syllable 2, e.g. *bu4fa3* 'illegal'; *traffic*) and half were s-words (with [s] in that position: *kuan1song1* 'loose'; *gossip*). Words were chosen such that using the other fricative would yield a nonword. The mean frequency for Mandarin words was 3.68 and 3.82 per million for f-words and s-words respectively (computed from the online CCL corpus of PKU [9]). The mean frequency for English words was 4.3 and 4.1 respectively for f-words and s-words (computed from SUBTLEX using the Zipf scale [10]).

For each language two versions of each training item were selected: one unaltered, and another with the critical word-medial fricative replaced by an ambiguous sound [?] (step 17 of the /fu/-/su/ continuum from the pre-test). 100 nonwords were created (in Mandarin this was by changing the tone of the second syllable in a real word (e.g., *ji1-dan4* 'egg' became *ji1-dan2*). Nonwords and fillers did not contain [f], [s], or [ʂ], [ɕ], [ts], or [tʃ] (to avoid perceptually similar sounds to the critical fricatives). The training materials were produced by the same speaker as for the [f]-[s] continuum.

3.1.3. Procedure

Participants completed two full sessions (lexically-guided perceptual learning followed by test), one in L1 and one in L2. These sessions were spaced 2-3 weeks apart. For each language, participants were first exposed to an ambiguous sound [?] either in f-words or s-words in a lexical decision task. They had to decide whether each item was a real word or a non-word, indicating their response via a button press, with "Yes" responses made using the dominant hand. Four stimulus lists were constructed, each containing the same 100 words and 100 non-words. Items were presented randomly with the restriction that no more than four words or non-words occurred in sequence. Two versions of each presentation order were created, one in which [?] replaced all instances of [f] ([f]-ambiguous group) and one in which [?] replaced all instances of [s] ([s]-ambiguous group); half of the participants were assigned to each group. The first 12 trials contained no instances of [?] and were identical across all versions and lists.

Following the lexical decision task, participants completed a categorisation task, in which they heard recordings of steps 7, 13, 17, 21, and 27 of the /fu/-/su/ continuum (identical steps were used for the Mandarin and English continua) and had to categorize each item as either /fu/ or /su/. These five steps were each presented randomly 30 times (150 trials in total).

For each categorisation task, perceptual learning was examined via a 2 × 5 ANOVA with the between-subjects factor of training group ([f]-ambiguous versus [s]-ambiguous) and the within-subjects factor of step (7, 13, 17, 21, 27).

3.2. Results and discussion

3.2.1. Mandarin categorisation

The analysis of the Mandarin-English late bilinguals' perceptual learning of Mandarin, their L1, showed as expected a significant main effect of group, $F(1, 20) = 6.2$, $p = .022$, $\eta_p^2 = .235$ (Figure 2). There was a significant main effect of step, $F(1.9, 38.7) = 199.0$, $p < .001$, $\eta_p^2 = .909$. There was a significant Group × Step interaction, $F(1.9, 38.7) = 6.1$, $p = .005$, $\eta_p^2 = .234$. We examined the interaction via a series of *t*-tests. Training groups differed on steps 13, $t(15.5) = 2.7$, $p = .016$, and 17, $t(20) = 3.5$, $p = .002$.

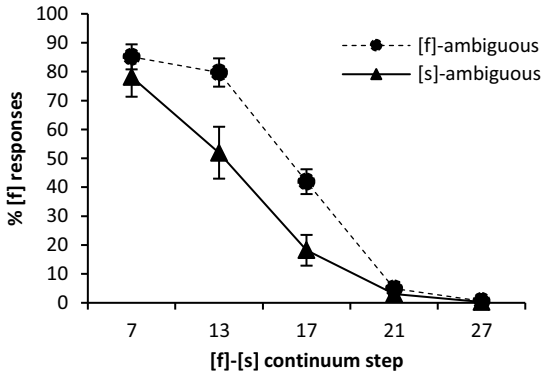


Figure 2: *Total proportion of [f] responses to a Mandarin [fu]-[su] continuum made by Mandarin-English late bilinguals following [f]-ambiguous or [s]-ambiguous training.*

3.2.2. *English categorisation*

The Mandarin-English late bilinguals' perceptual learning in English showed, in contrast to the L1 results, no significant main effect of group, $F(1, 20) = 0.4$, $p = .541$, $\eta_p^2 = .019$ (see Figure 3). There was a significant main effect of step, $F(2.5, 49.9) = 91.0$, $p < .001$, $\eta_p^2 = .820$, but no significant Group × Step interaction, $F(2.5, 49.9) = 0.6$, $p = .671$, $\eta_p^2 = .029$.

Thus for these listeners, their L1 performance replicated that found in previous work with Mandarin, but in their L2 they showed no evidence of successful perceptual learning.

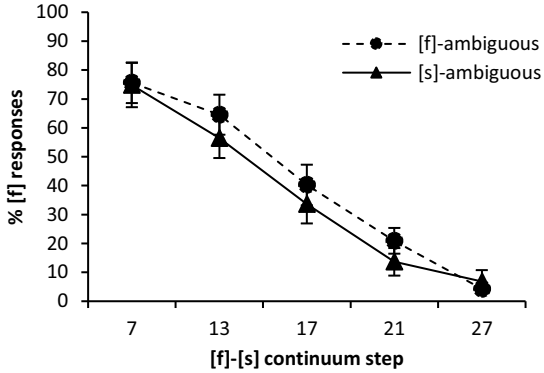


Figure 3: *Total proportion of [f] responses to an English [fu]-[su] continuum made by Mandarin-English late bilinguals following [f]-ambiguous or [s]-ambiguous training.*

4. Experiment 2

4.1. Method

4.1.1. *Participants*

Experiment 2 involved 25 Australian learners of Mandarin, again paid for participating. These participants were all born in Australia, had a mean age of 30.1 (range = 19.0-54.8), and had acquired Mandarin from a mean age of 14.9 (range = 6-29). All had Australian English as their dominant language. None reported any vision, hearing, or language impairments.

4.1.2. *Stimuli and Procedure*

These were as in Experiment 1.

4.2. Results and discussion

4.2.1. *English categorisation*

The Australian Mandarin learners' perceptual learning in their L1 English again showed the expected significant main effect of group, $F(1, 23) = 8.7$, $p = .007$, $\eta_p^2 = .274$ (see Figure 4). There was also a significant main effect of step, $F(2.7, 62.5) = 103.1$, $p < .001$, $\eta_p^2 = .818$, but there was no significant Group × Step interaction, $F(2.7, 62.5) = 1.5$, $p = .222$, $\eta_p^2 = .062$. We examined the group difference via a series of *t*-tests. Training groups differed on step 13, $t(23) = 2.2$, $p = .038$, step 17, $t(23) = 2.5$, $p = .018$, and step 21, $t(13.4) = 3.0$, $p = .010$. This result thus again replicates the perceptual learning effect in English, and replicates the significant learning observed so far in all experiments in an L1 in the L1 environment.

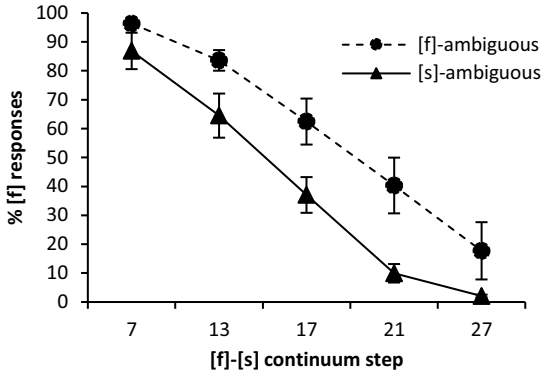


Figure 4: *Total proportion of [f] responses to an English [fu]-[su] continuum made by Australian learners of Mandarin following [f]-ambiguous or [s]-ambiguous training.*

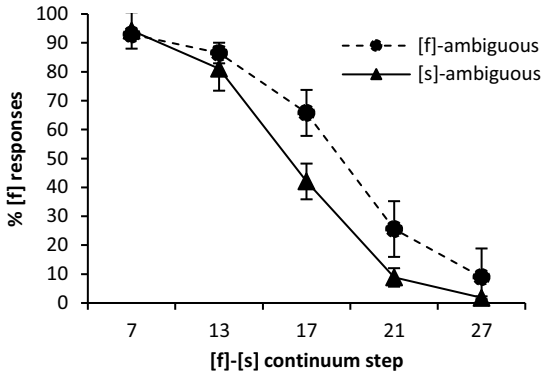


Figure 5: *Total proportion of [f] responses to a Mandarin [fu]-[su] continuum made by Australian learners of Mandarin following [f]-ambiguous or [s]-ambiguous training.*

4.2.2. *Mandarin categorisation*

The Australian Mandarin learners' perceptual learning in their L2 Mandarin showed a main effect of group that was marginally significant, $F(1, 23) = 3.3$, $p = .081$, $\eta_p^2 = .127$. There was a significant main effect of step, $F(2.9, 66.2) = 178.8$, $p < .001$, $\eta_p^2 = .886$. There was a significant Group × Step interaction, $F(2.9, 66.2) = 2.9$, $p = .046$, $\eta_p^2 = .110$. We examined the interaction via a series of *t*-tests. Training groups differed on step 17, $t(23) = 2.6$, $p = .017$, and the difference for step 21 was marginally significant, $t(12.8) = 1.824$, $p = .092$. Figure 5 shows these results.

Here again the results were weaker for these listeners' L2 than for their L1. However, unlike the Experiment 1 group, this learner population did show evidence of successful adaptation in their L2; the ambiguous sound that they had been trained on was categorised differently depending on their training (though they were not able to generalise the learning as widely across the phonemic continuum as they had done in their L1). The two learner groups that we have tested thus produced differing result patterns, with this difference being furthermore in the opposite direction from what previous L2 perceptual learning findings in the literature might have predicted: an immersion group was here less successful.

5. General Discussion

Our two experiments have revealed an asymmetry in the degree of perceptual learning achieved by L2 listeners within the language pair English-Mandarin Chinese. As expected from extensive prior investigations in other languages, both sets of listeners were able to adjust the [f]/[s] category boundary within their native language to adapt to an apparent talker idiosyncrasy in one of those phonemes. However in their L2 they were less successful, with the Mandarin listeners to English as their L2 performing worse than the Australian English listeners to Mandarin as their L2.

The dissimilarity between these two languages has allowed us to rule out speculations that perceptual learning would only be observed in L2 in the case of related L1/ L2 pairs. English and Mandarin are from different, quite unrelated, language families, and also differ on many independent phonological dimensions of relevance to the listening task. Mandarin has fewer vowels and fewer consonants than English, and it has simpler syllable structure and uses no morphological affixes. In all these respects its phonology is less complex than that of English. In the suprasegmental domain, however, Mandarin also has lexical tones that distinguish words, in which respect its phonology may be held to be more complex than English.

Usefully, with this pair we were able to exploit the same phonemic comparison ([f] versus [s]) as tested in many previous demonstrations of perceptual learning, from [2] on. Fricative perception typically leads to less markedly categorical functions than are seen with other consonants [11], but this is stable across languages. Fricative perception does differ across languages due to fricative inventory size and composition [12], but both Mandarin and English have, by world standards, relatively large fricative repertoires, albeit with somewhat greater competition for [f] in English and for [s] in Mandarin. Our results suggest that language similarity is not a prerequisite for the appearance of L2 perceptual learning.

Also, our study has provided new evidence on the role of linguistic immersion. Previous research showing perceptual learning in L2 had mostly been carried out in situations where the listeners were currently living in an environment in which their L2 was the expected language. Thus German students in the Netherlands showed a perceptual learning effect with Dutch input that was equivalent to that shown by L1 Dutch-speakers hearing the same training materials [5]. Dutch-born emigres in Australia showed perceptual learning in their L2 English (which for many of them had become their dominant language) to a more significant degree than in their original L1, Dutch [6]. Not only were both of these results drawing on a situation of immersion in the L2 linguistic environment, but they also involved related languages which are phonologically quite similar: Dutch, German, English. With the same similar languages, perceptual learning has also been reported with no immersion; [13] for German-English, [14] for Dutch-English.

However, given the fact that the listeners in our study who showed less evidence of perceptual learning (Experiment 1) were those in an immersion environment, while the more successful learners were those without immersion (Experiment 2), we no longer regard immersion as the most crucial factor determining the outcome of such studies. Instead, we suggest that future research should concentrate on determining the exact role of phonological asymmetries in the L1:L2 comparison. Although our stimuli were matched in both phonologies (with the critical phoneme in a word-medial syllable onset in each case), the two languages were, as described above, equally carefully chosen for their lack of phonological match.

Two testable hypotheses seem to merit attention regarding transfer to a second language of one's native facility with adaptation to new talkers via perceptual learning (recall the widespread success in this task across L1 investigations). Either of them could be responsible for the pattern of results we have found. First, it may be the case that transfer of this skill to a new phonological system is easier if the new system is simpler than that of the L1 (as in many ways the Mandarin system is simpler than the English system, given that English has more complex syllables, with both onset and coda clusters, plus morphological affixes on words, etc.). Alternatively, it may be the case that transfer is harder if the new system lacks certain elements that are crucial features of the L1 (as is the case, for instance, with English's lack of tones, which in Mandarin have been shown to be a suitable ground for perceptual learning [3]), or if the new system just has a larger phonemic inventory than the L1 (as English has, in comparison to Mandarin). Each of these hypotheses is testable using further language pairs.

6. Acknowledgements

This study was funded by the Australian Research Council (DP140104389, with additional support from the ARC Centre of Excellence in the Dynamics of Language; CE140100041).

7. References

[1] Cutler, A, Eisner, F, McQueen, J. M. and Norris, D. "How abstract phonemic categories are necessary for coping with speaker-related variation", in Papers in Laboratory Phonology, vol. 10, C. Fougeron, B. Kühnert, M.P. d'Imperio and N. Vallée, Eds. Berlin: Mouton de Gruyter, 2010, pp. 91-111.

[2] Norris, D. McQueen J.M. and Cutler A. "Perceptual learning in speech", Cognit. Psychol., 47: 204-238, 2003.

[3] Mitterer H. Chen, Y. and Zhou X., "Phonological abstraction in processing lexical-tone variation: Evidence from a learning paradigm", Cognit. Sci., 35: 184-197, 2010.

[4] Burchfield, L.A., Luk, S-H. K., Antoniou, M. and Cutler, A. "Lexically guided perceptual learning in Mandarin Chinese", Proc. Interspeech 2017, Stockholm, pp. 575-580, 2017.

[5] Reinisch, E., Weber, A. and Mitterer, H. "Listeners retune phoneme categories across languages". J. Exp. Psychol: Hum Percept. Perform., 39: 75-86, 2013.

[6] Bruggeman, L. "Nativeness, dominance, and the flexibility of listening to spoken language." PhD thesis, Western Sydney University, 2016.

[7] McQueen, J. M. Tyler, M. D. and Cutler, A. "Lexical retuning of children's speech perception: Evidence for knowledge about words' component sounds." Lang. Learn. Dev., 8: 317-339, 2012.

[8] Boersma, P. and Weenink, D. "Praat: Doing phonetics by computer" (Version 6.0.27) [Computer program; Available: http://www.praat.org/], 2017.

[9] Center for Chinese Linguistics PKU. 语料库检索系统 (网络版) http://ccl.pku.edu.cn:8080/ccl_corpus/ [Online]. Accessed March 19, 2017.

[10] Van Heuven, W. J. B., Mandera, P., Keuleers, E. and Brysbaert, M. "Subtlex-UK: A new and improved word frequency database for British English". Quart. J. Exp. Psychol., 67:1176-1190, 2014.

[11] Repp, B.H. "Two strategies in fricative discrimination", Perc. Psychophys., 30: 217-227, 1981.

[12] Wagner, A., Ernestus, M. and Cutler, A. "Formant transitions in fricative identification: The role of native fricative inventory". J. Acoust. Soc. Am., 120: 2267-2277. 2006.

[13] Schuhmann, K.S. "Perceptual learning in second language learners." PhD thesis, Stony Brook University, 2014.

[14] Drozdova, P., Van Hout, R. and Scharenborg, O. "Lexically guided perceptual learning in non-native listening", Biling. Lang. Cogn., 19: 914-920, 2016.

The Effects of Foreign Language Learning on the Perception of Japanese Consonant Length Contrasts

Kimiko Tsukada[1], *Kaori Idemaru*[2], *John Hajek*[3]
[1]Macquarie University, Australia
[2]University of Oregon, USA
[3]The University of Melbourne, Australia
kimiko.tsukada@gmail.com, idemaru@uoregon.edu, j.hajek@unimelb.edu.au

Abstract

The perception of Japanese singleton/geminate contrasts by native and non-native listeners was compared to examine if not only specific but general foreign language experience might facilitate the processing of unfamiliar sounds. Three groups of non-native listeners and a native Japanese control group participated in the AXB discrimination experiment. As expected, native Japanese listeners outperformed non-native listeners in discriminating Japanese length contrasts. While learners of Japanese did not match the native level of performance, they were more accurate than listeners with other foreign language experience, suggesting that general experience may not transfer positively to the processing of Japanese length contrasts.

Index Terms: cross-language speech perception, consonant length contrasts, Japanese, singleton/geminate

1. Introduction

Japanese uses length (e.g. short vs long) contrastively for both vowels and consonants [1]. For example, /kite/ *'wearing'* contrasts with /ki:te/ *'listening'* on the one hand and with /kit:e/ *'stamp'* on the other hand. Duration is the primary (though not the only) acoustic cue to differentiate the short and long members of the length contrast [2, 3]. It is widely acknowledged that length contrast is difficult to learn for non-native learners from diverse first language (L1) backgrounds [4-9]. Length contrasts may be difficult to acquire, because they are not as frequent cross-linguistically or as robust as other phonetic contrasts such as the voicing contrast (e.g. *tip* vs. *dip*), which is supported by multiple co-varying acoustic cues [10].

This study examined the role of foreign language (FL) learning experience in the processing of unfamiliar phonological contrasts. We asked native speakers of American English who were enrolled in Japanese or Spanish courses at a university to discriminate Japanese consonant length contrasts in an AXB task. Spanish was chosen, as it is one of the most popular FLs in the US. However, unlike Japanese, neither American English nor Spanish uses consonant length contrastively within a word. We were interested in examining if not only listeners with experience in Japanese but also alternatively in Spanish, an unrelated language, may outperform monolingual listeners from the same L1 background (i.e. American English).

We acknowledge that experience in Spanish may not provide specific advantage in the processing of Japanese singleton/geminate contrasts. This is because, as mentioned above, Spanish does not use duration for lexically contrastive purposes [11]. However, learning any FL may be generally facilitative, as it requires learners to pay attention to phonetic/phonological features absent in their L1 [12]. By comparing learners of not only Japanese but learners of unrelated language (Spanish) to monolingual American English speakers, we hope to gain an insight into the role of specific vs general FL learning. A group of 10 native Japanese (NJ) speakers was included as controls. We would expect NJ listeners to outperform non-native listeners, as they should have firm cognitive representations for singleton/geminate contrasts in their L1 Japanese. The non-native learners of Japanese (JPN) would also be expected to outperform the other two groups of non-native listeners, as the former has experience learning FL Japanese and should have greater awareness of Japanese length categories than those without knowledge of Japanese. It is unclear if learners of Spanish (SPAN) might outperform native English (NE) speakers without extensive FL experience.

Research on FL speech learning focusing on non-segmental features such as length is increasing but still relatively limited, in particular, for target languages other than English. The knowledge gained from the present research will advance our current understanding of not only L1 but also FL transfer effects in cross-language speech perception and language learning in adulthood.

2. Methods

The purpose of this experiment was to compare the perception of Japanese consonant length by native and non-native listeners with varying linguistic experience. It examined the discrimination accuracy of singleton/geminate contrasts via a forced-choice AXB discrimination test. The stimuli were produced by multiple NJ speakers as described below.

2.1. Speakers, stimuli and procedure

Six NJ speakers (3 males, 3 females) participated in the recording sessions, which lasted between 45 and 60 minutes. The speakers' age ranged from late twenties to early forties. According to self-report, which was confirmed by the first author who is a NJ speaker originally from Tokyo, all NJ speakers spoke standard Japanese, having been born or having spent most of their life in the Kanto region surrounding the Greater Tokyo Area. The NJ speakers were recorded in the recording studio at the National Institute of Japanese Language and Linguistics, Tokyo.

Table 1 shows 12 Japanese word pairs used in this study. The /(C)VC(C)V/ tokens contained singleton or geminate consonants intervocalically. As the accent type (High-Low (HL) and LH) has been reported to affect native and non-native listeners' length perception in Japanese [6, 13-15], tokens with both HL and LH were included in the stimuli. Only tokens with

stops were considered in this study. As voiced geminates are very limited in Japanese, only voiceless stops (/t k/) were used. To record the stimuli, each word was presented on a computer screen in random order and produced in two separate conditions: one in isolation and the other in a carrier sentence (/sokowa _____ to jomimasu/ *'You read it as _____ there'*). The pace of presentation was controlled by the experimenter (the first author). The speech materials were digitally recorded at a sampling rate of 44.1 kHz and the target words were segmented and stored in separate files. To avoid inter-speaker variation in fluency (specifically, the duration of a pause before and after the target word), only tokens produced in isolation were used as experimental stimuli in this study.

Table 1: *Japanese words and non-words used.*

	Without geminate		With geminate	
consonant	HL	LH	HL	LH
/t/	/kato/ *transition*	/heta/ *unskilled*	/kaːto/ *cut*	/heːta/ *decreased*
	/mate/ *wait*	/oto/ *sound*	/maːte/ *waiting*	/oːto/ *husband*
	/sate/ *well, then*	/wata/ *cotton*	/saːte/ *leaving*	/waːta/ *broke*
/k/	/ika/ *below*	/ake/ *open*	/ikːa/ *lesson one*	/akːe/ *appalled*
	/kako/ *past*	/haka/ *grave*	/kakːo/ *parenthesis*	/hakːa/ *mint*
		/saka/ *slope*		/sakːa/ *author*
		/ʃike/ *rough sea*		/ʃikːe/ *humidity*

The format of the perception experiment was a forced-choice AXB discrimination test. The presentation of the stimuli and the collection of perception data were controlled by the PRAAT program [16]. In the AXB test, the first (A) and third (B) tokens always came from different length categories, and the listeners 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 listeners listened to 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 listeners focused on relevant phonetic characteristics that group two tokens as members of the same category without being distracted by audible but phonetically irrelevant within-category variation (e.g. in voice quality). This was considered a reasonable measure of listeners' perceptual capabilities in real world situations [17]. All possible AB combinations (i.e. AAB, ABB, BAA, and BBA 48 trials each) were tested.

The listeners were given two ('A', 'B') response choices on the computer screen. They were asked to click on the option 'A' if they thought that the first two tokens in the AXB sequence were the same and to click on the option 'B' if they thought that the last two tokens were the same. No feedback was provided during the experimental sessions. The listeners could take a break after 50 trials if they wished. The listeners 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 listener wished, but responses could not be changed once given. The inter-stimulus interval in all trials was 0.5 s.

2.2. Participants

Four groups of listeners participated. The first group consisted of 14 (6 males, 8 females, *mean age* = 23.3 years, *sd* = 4.2) native speakers of American English who were learners of Japanese (JPN) at different levels of proficiency. One of them was a heritage learner who was enrolled in the introductory level Japanese. Excluding this heritage learner, the JPN listeners started learning Japanese at the age of 18.2 on average (*sd* = 6.1) and had a mean length of learning of 3.7 (*sd* = 4.0) years. The second group consisted of 10 (4 males, 6 females, *mean age* = 23.3 years, *sd* = 3.8) native speakers of American English who were learners of Spanish (SPAN) at different levels of proficiency. One of them was a heritage learner. Excluding this heritage learner, the SPAN listeners started learning Spanish at the age of 14.3 on average (*sd* = 2.7) and had a mean length of learning of 5.6 years (*sd* = 1.8). The third group consisted of 10 (3 males, 7 females, *mean age* = 19.4 years, *sd* = 0.8) native speakers of American English (NE) who were enrolled in Psychology or Linguistics courses and received credit for research participation. None of the NE listeners had experience learning Japanese or Spanish formally at college level, but this is not intended to guarantee their monolingualism and participants' language background needs to be more tightly controlled in future work. The fourth and last group consisted of 10 (2 males, 8 females, *mean age* = 21.0 years, *sd* = 0.8) NJ listeners who participated as control subjects. All of the NJ listeners were born and spent the majority of their life in Japan. The NJ listeners started learning English at the age of 11.1 on average (*sd* = 2.3). Their mean length of residence in the US was 0.4 years (*sd* = 0.22). None of them participated in the recording sessions. According to self-report, all four groups of listeners had normal hearing.

All listeners were tested individually in a session lasting approximately 30 to 40 minutes in a sound-attenuated laboratory at a university in USA. The experimental session was self-paced. They heard the stimuli at a self-selected, comfortable amplitude level over the high-quality headphones (AKG K240 MKII) on a desktop computer in the sound-treated experiment room.

3. Results

Table 2 shows the mean discrimination scores (%) for each group of listeners broken down to the trials where the target stop was either singleton or geminate. Although the JPN and SPAN groups included language learners at different levels of proficiency and were not as homogeneous as we would have liked, the discrimination scores were averaged, as the number of listeners in each group was relatively small. Data collection is still in progress.

The mean discrimination accuracy ranged from 74.3 (*sd* = 12.2) for the NE group to 98.8 (*sd* = 1.3) for the NJ group. The JPN group discriminated the length contrasts with greater than 90% accuracy. This is considerably higher than the NE group and demonstrates a positive effect of L2 Japanese speech learning. The number of JPN listeners whose discrimination scores reached the range set by the NJ group was 2 (14%) for target singleton and 10 (71%) for target geminate, respectively. Thus, the JPN group was more accurate when the target token presented in the X position contained a geminate than when it contained a singleton.

Table 2: *Mean discrimination scores (%) by four groups of listeners. Standard deviations are in parentheses.*

Group	Target singleton	Target geminate
NJ (n = 10)	99.0 (1.4)	98.8 (2.1)
JPN (n = 14)	91.3 (5.6)	94.1 (4.2)
SPAN (n = 10)	78.7 (8.8)	84.7 (9.0)
NE (n = 10)	73.4 (10.9)	75.6 (14.8)

Figure 1 shows the distribution of discrimination accuracy scores as a function of listener group and length category of the target stop. It can be clearly seen that the NJ group was highly homogeneous and accurate in discriminating singleton and geminate stops in their L1 Japanese.

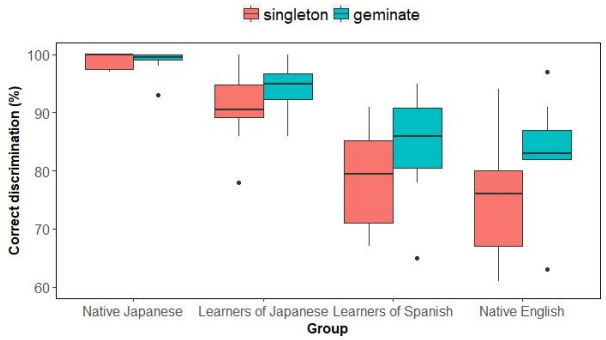


Figure 1: *The distribution of discrimination scores (%) for trials with the target singleton or geminate by four groups of listeners. The horizontal line in the box indicates the median. The bottom and top lines of the box indicate the first and third quartiles. The points outside the box are outliers.*

Table 3: *Results of Group x Length ANOVA.*

Factor	*df*	*F*-value	*p*-value	η^2_G
G	3, 40	21.6	< .001	.59
L	1, 40	10.9	< .01	.03

Table 4: *Results of one-way ANOVA assessing the effects of Group and multiple comparison tests (significance level at .05).*

df	*F*-value	*p*-value	estimated ω^2	Between-group comparisons
3, 17.9	30.6	< .05	.67	NJ > JPN, SPAN, NE, JPN > SPAN, NE

Table 3 shows the results of two-way (Group x Length) analysis of variance (ANOVA). The G × L interaction was not significant, which suggests that the pattern of singleton vs geminate discrimination was comparable across the four groups of listeners. In other words, length contrasts were more discriminable when the target stop was a geminate than when it was a singleton. Table 4 shows the results of one-way ANOVA which assessed the effect of Group (not assuming equal variances) and Dunnett's Modified Tukey-Kramer pairwise multiple comparison *post hoc* tests. It appears that learning Spanish did not give the SPAN listeners advantage over the NE listeners in processing length contrasts in Japanese.

We also conducted a preliminary analysis to examine the effect of pitch accent type on the discrimination of consonant length. Figure 2 shows the listeners discriminated consonant length contrasts more accurately when the pitch accent was HL than when it was LH. Table 5 shows the results of two-way (Group x Pitch) ANOVA. While between-group differences are clearly visible for both accent types in Figure 2 [HL: $F(3, 17.9) = 18.3$, $p < .001$, LH: $F(3, 18.0) = 36.8$, $p < .001$], the simple effect of Pitch was significant only for the JPN group [$F(1, 25.1) = 7.2$, $p < .05$, estimated $\omega^2 = .18$].

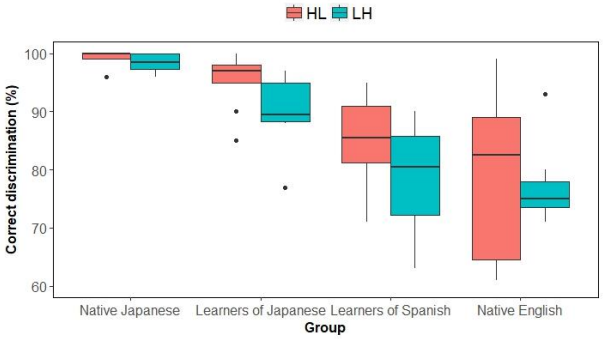


Figure 2: *The distribution of discrimination scores (%) for trials with HL or LH pitch accents by four groups of listeners. The horizontal line in the box indicates the median. The bottom and top lines of the box indicate the first and third quartiles. The points outside the box are outliers.*

Table 5: *Results of Group x Pitch ANOVA.*

Factor	*df*	*F*-value	*p*-value	η^2_G
G	3, 40	20.7	< .001	.59
P	1, 40	61.2	< .001	.10
G x P	3, 40	5.0	< .01	.03

4. Discussion

This study examined discrimination accuracy of Japanese consonant length by native and non-native listeners differing in their FL experience. Three groups of non-native listeners and a control group of NJ listeners participated. Two groups of non-native listeners had FL experience in Japanese or Spanish, a language unrelated to Japanese. The third group of non-native listeners had no college-level FL experience.

As expected, the NJ listeners were highly accurate and outperformed the non-native listeners. The JPN listeners were more accurate than both the SPAN and NE listeners who were comparable in their discrimination of unfamiliar Japanese length contrasts. Although the JPN listeners were less accurate than the NJ listeners in discriminating length contrasts, the majority (71%) of the JPN listeners' scores reached the range set by the NJ group when the target stop was a geminate. Thus, it is unlikely that there is an absolute limit in non-native learners' ability to achieve truly native-like length perception. Although the SPAN listeners' mean discrimination scores were higher than that of the NE listeners (78.7% vs 73.4% for the target singleton and 84.7% vs 75.6% for the target geminate), the between-group difference did not reach significance, possibly due to large individual variation as seen in Figure 1. As the number of participants in each group is relatively small, these results need to be interpreted with caution.

We observed that listeners' length discrimination accuracy was higher for the trials with tokens containing a geminate in the target position than those containing a singleton. It is not entirely clear what made the length contrasts more discriminable when the target stop was a geminate than when it was a singleton. This length effect may be related to the task used in this study. In the AXB discrimination test, listeners need to focus on the token in the middle (X) position and decide whether it belonged to the same category as the first token (A) or the third token (B). Possibly, stop closure duration for singletons was too short (i.e. below just noticeable difference in duration) especially for non-native listeners to process accurately. Further, unlike many languages that use length phonemically [18, 19], in Japanese, vowels tend to be phonetically *longer* before geminate obstruents than before their single counterparts [4, 20]. This type of cross-linguistic phonetic characteristics may affect listeners' length perception. Another possibility is that the germinates, which do not occur in English or Spanish, simply stand out more as they are unfamiliar features to the listeners, and hence easier to spot.

Although the JPN listeners clearly outperformed the NE listeners without formal FL learning experience at the college level and resembled the NJ listeners to a greater extent than other non-native listeners, these non-native JPN learners may still need to expend more effort or allocate more attentional resources in discriminating length contrasts in Japanese. To examine whether JPN learners have developed cognitive representations for Japanese length categories that are expected of NJ listeners, increasing task difficulty (e.g. stimuli presented at different speaking rates) may be useful.

The SPAN listeners did not significantly differ from the NE listeners. Thus, general FL experience may be of limited help in cross-linguistic speech perception of unfamiliar languages. Spanish does not use length contrastively as was mentioned in the Introduction. On the other hand, English-speaking learners of languages such as Thai, Italian or Arabic that use length phonemically for vowels, consonants or both may be able to utilize their FL knowledge and may be significantly more accurate in processing unfamiliar Japanese consonant length than the NE listeners.

To determine if FL learners of Japanese ever attain genuinely native-like length perception and, if so, how long it might take them to achieve that level of accuracy, it would be necessary to include JPN learners with even more experience in Japanese. It would also be useful to conduct a longitudinal study to gain a better understanding of the time course of learners' speech development.

5. Conclusions

American English-speaking learners of Japanese were significantly more accurate in discriminating Japanese consonant length contrasts than naïve listeners from the same L1 background, demonstrating a positive effect of FL experience on the perception of non-native contrasts. To the extent that learners of Spanish did not differ from listeners without FL experience in discriminating unknown Japanese length contrasts, general FL learning experience may not transfer positively to the processing of speech sounds in an unfamiliar language.

6. Acknowledgements

This research was supported by the 2018 Endeavour Research Fellowship. We thank faculty members of the Departments of East Asian Languages & Literatures and Romance Languages, University of Oregon for help with participant recruitment, Hayli Brown and Laura Walter for research assistance and participants for making the study possible. Finally, we thank the three anonymous reviewers for their time and input.

7. References

[1] Vance, T., The Sounds of Japanese, Cambridge University Press, 2008.

[2] Fujisaki, H., Nakamura, K. and Imoto, T. "Auditory perception of duration of speech and nonspeech stimuli", Annual Bulletin, Research Institute of Logopedics and Phoniatrics, 7: 45-64, 1973.

[3] Kawahara, S. "The phonetics of sokuon, or geminate obstruents", in Kubozono, H [Ed], Handbook of Japanese Phonetics and Phonology, 43-78, De Gruyter Mouton, 2015.

[4] 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.

[5] Harada, T. "The acquisition of single and geminate stops by English-speaking children in a Japanese immersion program", Studies in Second Language Acquisition, 28: 601-632, 2006.

[6] Hung, H. Y. "Perception of Japanese geminate stops among Taiwanese learners of Japanese", Journal of the Phonetic Society of Japan, 16: 15-27, 2012.

[7] Kubozono, H. "Introduction to the special issue on Japanese geminate obstruents", Journal of East Asian Linguistics, 22: 303-306, 2013.

[8] Sonu, M., Kato, H., Tajima, K., Akahane-Yamada, R. and Sagisaka, Y. "Non-native perception and learning of the phonemic length contrast in spoken Japanese: Training Korean listeners using words with geminate and singleton phonemes. Journal of East Asian Linguistics, 22: 373-398, 2013.

[9] Toda, T. "Issues regarding geminate consonants in Japanese language education", Journal of the Phonetic Society of Japan, 11: 35-46, 2007.

[10] Lisker, L. "'Voicing' in English: A catalogue of acoustic features signaling /b/ versus /p/ in trochees", Language and Speech, 29: 3-11, 1986.

[11] Martínez-Celdrán, E., Fernández-Planas, A. M. and Carrera-Sabaté, J. "Illustrations of the IPA: Castilian Spanish", Journal of the International Phonetic Association, 33: 255-260, 2003.

[12] Cenoz, J. "The influence of bilingualism on third language acquisition: Focus on multilingualism", Language Teaching, 46: 71-86, 2013.

[13] Minagawa, Y. and Kiritani, S. "Discrimination of the single and geminate stop contrast in Japanese by five different language groups", Annual Bulletin, Research Institute of Logopedics and Phoniatrics, 30: 23-28, 1996.

[14] Ofuka, E. "Perception of a Japanese geminate stop /tt/: The effect of pitch type and acoustic characteristics of preceding/following vowels", Journal of the Phonetic Society of Japan, 7: 70-76, 2003.

[15] Tsukada, K., Cox, F., Hajek, J. and Hirata, Y. "Non-native Japanese learners' perception of consonant length in Japanese and Italian", Second Language Research, 34: 179-200, 2018.

[16] 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).

[17] 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.

[18] Maddieson, I. "Phonetic cues to syllabification", in V. Fromkin [Ed], Phonetic Linguistics: Essays in Honor of Peter Ladefoged, 203-221, Academic Press, 1985.

[19] Hajek, J., Stevens, M., and Webster, G. "Vowel duration, compression and lengthening in stressed syllables in Italian", Proceedings of the 16th International Congress of Phonetic Sciences: 1057-1060, 2007.

[20] Idemaru, K. and Guion, S. "Acoustic covariance of length contrast in Japanese stops", Journal of the International Phonetic Association, 38: 167-186, 2008.

Exploring sub-band cepstral distances for more robust speaker classification

Takashi Osanai[1], Yuko Kinoshita[2], Frantz Clermont[3]

[1] National Research Institute of Police Science, Japan
[2]College of Arts and Social Science/Asia and the Pacific, The Australian National University
[3]J.P. French Associates Forensic Lab., England
osanai@nrips.go.jp; yuko.kinoshita@anu.edu.au; dr.fclermont@gmail.com

Abstract

This paper presents the first of two-part exploration into the potential of parametric cepstral distance (PCD) as a forensic voice comparison feature, based on Japanese vowel data collected from 306 male native speakers under microphone and mobile transmission conditions. The behaviours of PCDs were closely examined by altering sub-band settings, and we found the behaviour of PCDs to correspond well to what is known about formants, which suggests that PCDs are relatable to articulatory gestures. Comparison between sub-band and full-band PCD revealed that limiting the band range to a specific frequency region makes the feature more robust against channel mismatch, encouraging further examination of this potential feature.

Index Terms: Sub-band cepstral distance, F-ratio, Speaker Classification, Channel mismatch, Japanese vowels.

1. Introduction

In the past few decades, the field of forensic voice comparison (FVC henceforth) has seen considerable development in its methodology in classification techniques and in the evaluation of the systems themselves (e.g. [1-4]). Still, the features to which such techniques are applied are mostly unchanged: formants and various types of cepstra, such as MFCC or LPCC, appear to be the two most commonly used features. Some past FVC research favours the use of cepstrum-based features; they generally outperform formants in speaker classification (e.g. [5-7]). This is unsurprising given the differences in their nature as features. Formants represent only the locations of spectral peaks in the frequency domain, whereas the cepstrum captures richer information by utilising the entire user-defined frequency range. Also, the cepstrum can be extracted automatically. Automatic formant extraction, on the other hand, is known to be highly unreliable (e.g. [8]) and often requires manual supervision and correction. This leads to two problems: introducing measurer-dependent variability to the data (e.g. [8-10]), and extreme resource intensiveness.

However, formants have two major advantages over the cepstrum: robustness and interpretability. A real-life FVC case often involves data of poor quality, recorded through different devices and transmission channels. Formants are known to be more robust than the cepstrum under such conditions. Formant frequencies also generally correspond to articulatory gestures in speech production, and it is therefore easier to communicate their meaning to the layperson, as well as for the expert to detect any unusual characteristics or irregularity in the data, which may or may not be related to speaker characteristics.

In legal proceedings, experts are tasked to assist the court to reach correct decisions. Communicating their analysis processes and outcomes in an understandable way to non-experts is thus essential. Of course discussing scientific evidence inevitably involves highly technical concepts unfamiliar to laypeople, and "what is understandable?" is arguable, but it is our view that less abstract and more intuitively understandable processes are preferable in these contexts.

Thus we believe that ideal FVC features need qualities additional to the standard requirements of discriminability: extracted automatically and reliably; robust against poor recording quality and unpredictable environments; and relatable to articulatory gestures for better interpretability.

The band-limited parametric cepstral distance (sub-band PCD) proposed in [11] was identified in [12] as a feature that potentially meets such criteria. Firstly, sub-band PCD is a cepstrum-based feature and readily extracted without human supervision. This facilitates large-scale data processing and excludes measurer-dependent variability. It also allows the analysis to exclude unwanted frequency ranges which largely carry non-speech information. This is particularly attractive in FVC contexts, as recordings in real life FVC often contain substantial background noise, such as passing cars, other peoples' voices, and television noise. The level and the characteristics of such noise sources vary from one moment to next, so the capacity to flexibly focus on relevant frequency ranges should be a significant advantage.

While the preliminary investigation by [12] was based on a very small dataset, it made a few promising observations. First, the F-ratios of sub-band PCDs appear to correspond well to those from formants. Also, the F-ratios were very similar across the mobile transmission and microphone recordings, suggesting that sub-band PCD may be robust against transmission mismatches.

These observations call for further investigation on this feature based on a much larger dataset. The current study thus presents the first part of this investigation. Focusing on observations to the channel effect in relation to the F-ratio and the effect of differing frequency ranges and regions of the sub-bands, we aim to better understand the behaviours of sub-band PCD and explore its potential as a feature for FVC in court.

2. Data

This study selected 306 adult male speakers from the NRIPS database [13]. They are native speakers of Japanese, aged from 18 to 76 years. Their places of origin spread widely across Japan, hence so did their dialectal background. Two non-contemporaneous recordings, separated by 2 to 3 months, were made for each speaker and the recording tasks were performed twice at each recording session. Recordings were made simultaneously through 2 channels: direct microphone (Ch1), and mobile phone transmission (Ch3).

The speech material consisted of read (C)V syllables: the 5 Japanese vowel phonemes, /a/, /i/, /ɯ/, /e/, and /o/, in combination with the 11 preceding consonants, ∅ (no consonant), /k/, /s/, /t/, /h/, /r/, /g/, /z/, /d/, /b/, and /p/. We excluded the consonants /n/, /m/, /y/, and /w/, for the in order to facilitate reliable automatic segmentation.

Japanese hiragana syllabary pairs, i.e. ぢ /di/ – じ /zi/ and づ /dɯ/ – ず /zɯ/, have been merged into phonetically identical forms, [dʐi] and [dzɯ], although the writing system still maintains the distinction. Therefore, for /i/ and /ɯ/, we have the vowel data in 10 different phonological contexts and 11 for /a/, /e/, and /o/.

3. Procedures

3.1. Segmentation and full-band LPCC extraction

The target syllables were automatically segmented into a preceding consonant and a vowel based on their power and F0. The sound files were down-sampled from 44.1 kHz to 8kHz, and full-band LPCCs were extracted from the selected vowel sections (order 14, Hamming window, window length 25ms, time-step 5ms). The LPCCs was averaged across the vowel duration, and the means across different phonological contexts were calculated for each vowel. As result, we obtained the LPCCs for 5 vowels, 2 recording sessions, 2 repeats, and 2 recording channels for each speaker.

3.2. Parametric cepstral distance and F-ratio calculation

The usefulness of FVC features is reflected in the ratio of between- to within-speaker variances, which are expressed below by Equations (1) and (2), respectively. The numerators of both expressions describe parametric cepstral distances (PCDs) between pairs of full-band LPCCs that are index-weighted by the matrix $\mathbf{K}$ to emphasise spectral slope differences, and weighted by the matrix $\mathbf{W}(\omega_1, \omega_2)$ to focus on any sub-band selectable by its lower and upper limits ω_1 and ω_2. It should be noted that the formulation of $\mathbf{W}$ detailed in [1] affords the flexibility of obtaining sub-band PCDs directly from full-band LPCCs. Note also that, for $\omega_1 = 0$ and $\omega_2 = \pi$, the distances in equations (1) and (2) simply reduce to full-band PCDs.

$$\sigma^2_{between}(\omega_1, \omega_2) = \frac{\sum_{i=1}^{N} n_i \, d_i^2(\omega_1, \omega_2)}{N - 1} \qquad (1)$$

$$\sigma^2_{within}(\omega_1, \omega_2) = \frac{\sum_{i=1}^{N} \sum_{j=1}^{n_i} d_{ij}^2(\omega_1, \omega_2)}{(\sum_{i=1}^{N} n_i) - N} \qquad (2)$$

where:
$i \equiv$ speaker-session index, $N \equiv$ number of speakers
$j \equiv$ token index, $n_i \equiv$ number of tokens per i^{th} speaker

$$d_i^2(\omega_1, \omega_2) = \left(\overline{\mathbf{C}_i} - \bar{\bar{\mathbf{C}}}\right)^T \cdot \mathbf{K}^T \cdot \mathbf{W}(\omega_1, \omega_2) \cdot \mathbf{K} \cdot \left(\overline{\mathbf{C}_i} - \bar{\bar{\mathbf{C}}}\right)$$
$$\equiv \text{PCD between } \overline{\mathbf{C}_i} \text{ and } \bar{\bar{\mathbf{C}}} \qquad (3)$$

$$d_{ij}^2(\omega_1, \omega_2) = \left(\mathbf{C}_{ij} - \overline{\mathbf{C}_i}\right)^T \cdot \mathbf{K}^T \cdot \mathbf{W}(\omega_1, \omega_2) \cdot \mathbf{K} \cdot \left(\mathbf{C}_{ij} - \overline{\mathbf{C}_i}\right)$$
$$\equiv \text{PCD between } \mathbf{C}_{ij} \text{ and } \overline{\mathbf{C}_i} \qquad (4)$$

$\mathbf{C}_{ij} \equiv$ mean LPCC for i^{th} speaker's j^{th} token across the vowel duration
$\overline{\mathbf{C}_i} \equiv$ mean LPCC for i^{th} speaker across all tokens
$\bar{\bar{\mathbf{C}}} \equiv$ grand-mean LPCC over all speakers

$$\mathbf{K} = \begin{bmatrix} \mathbf{1} & \mathbf{0} & \cdots & \cdots & \mathbf{0} \\ \mathbf{0} & \ddots & \ddots & & \vdots \\ \vdots & \ddots & \boldsymbol{k} & \ddots & \vdots \\ \vdots & & \ddots & \ddots & \mathbf{0} \\ \mathbf{0} & \cdots & \cdots & \mathbf{0} & \boldsymbol{M} \end{bmatrix} \equiv \text{index-weighting matrix}$$

$\mathbf{W}(\omega_1, \omega_2) \equiv$ band-selective matrix (see [1])
$\boldsymbol{M} \equiv$ LPCC order
$\omega_1 \equiv$ lower limit of sub-band selected within $[0, \pi]$
$\omega_2 \equiv$ upper limit of sub-band selected within $[0, \pi]$

To observe interactions between PCDs and sub-band ranges, we divided the range from 100 Hz to 1000 Hz in 100 Hz increments, and the range from 1000 Hz to 3500 Hz in 500-Hz increments. In search for the frequency regions that are relatively rich in speaker information, we also shifted sub-bands by 100-Hz steps, and calculated PCDs for each step.

3.3. Comparisons

The four recordings from 306 speakers yielded 1224 patterns of same-speaker (SS) pairs and 373,320 patterns of different-speaker (DS) pairs. The 2 different recording channels allowed us to make comparisons under 3 different channel conditions: Ch1 (microphone), Ch3 (mobile phone), and Ch1-3 (channel mismatch).

After calculating the sub-band and full-band PCDs, we examined their F-ratios and conducted simple verification tests based on the size of PCDs, so that we can observe the speaker-classification potential of each vowel-and-sub-band combination. PCDs are already a distance measure, so we pooled PCDs from 1224 SS pairs and 373,320 DS pairs, and separately plotted for their distributions. Using the Equal Error Rate (EER) as the threshold, we classified the PCDs under the three different channel conditions.

4. Results and discussion

4.1. F-ratios

We observed higher F-ratios for the between-Ch1 comparison than for between-Ch3, with occasional exceptions. This was expected, as microphone recordings contain more information and less of the unpredictable variability caused by mobile phone and telephone transmission.

The F-ratio plots for various sub-band ranges revealed that sub-band PCDs generally behave similarly in both channels, although how much they vary across two channels depends on vowel, with /a/ and /o/ revealing the greater variation than the rest. In general, the difference between two channels were greater in the higher-frequency regions.

Also, for all vowels, the greatest peak F-ratios were found where the sub-band range was set at 100Hz, the narrowest of all. The frequency region affects F-ratio less as the sub-band range becomes greater, and gets close to the baseline F-ratio, obtained from the full-band PCD, as expected.

In observing their relation to the full-band PCDs, we find that the sub-band PCDs outperform the full-band PCDs in certain frequency regions. This suggests that the sub-band PCDs are likely to outperform the full-band PCDs in speaker classification, when multiple of them are combined.

Figure 1 presents the results for one of the sub-band sizes, 300 Hz, as an example. It summarises the relationship between F-ratios and frequency regions. The results from Ch1 are shown in red, and Ch3 in blue. The horizontal lines present the baseline F-ratio that was produced form the full-band PCDs.

One of the characteristics we seek in FVC features are interpretability. If we can observe the correspondence between the expected formant values and the frequency regions which produced the high F-ratios, it suggests sub-band PCDs' interpretability similar to formants. Here we present Japanese vowel formants data obtained from 11 male speakers [14] to compare with Figure 1. It shows that the frequency regions where higher F-ratios was observed are, in many cases, vicinity of the expected locations of the formants.

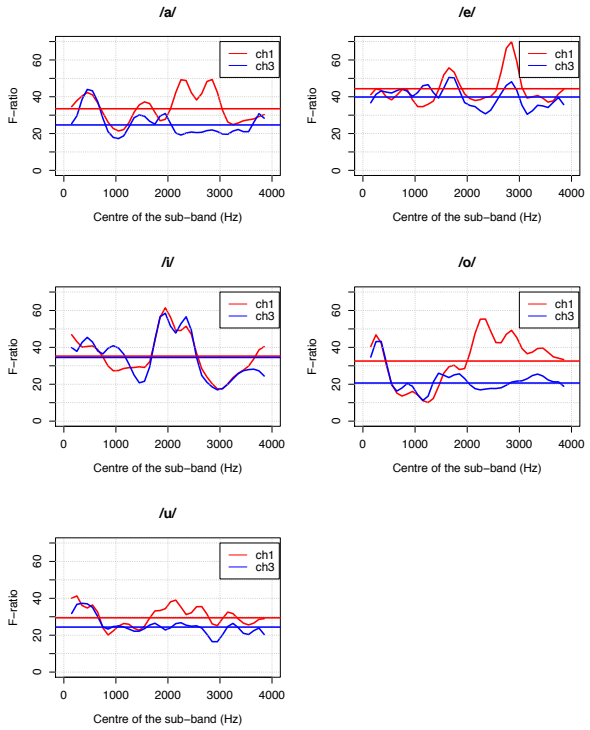


Figure 1: *F-ratios for varying frequency regions (sub-band width 300Hz shifted by 100Hz step), horizontal lines indicating full-band results*

Table 1: *Mean and SD of vowel formants from 11 male speakers of Japanese*

	F1		F2		F3	
	mean	*sd*	*mean*	*sd*	*mean*	*sd*
/a/	666	53	1414	114	2532	190
/e/	452	108	2009	136	2666	166
/i/	306	35	2174	184	2886	189
/o/	462	94	1125	115	2508	163
/u/	334	45	1550	154	2404	202

4.2. Verification results

Next we conducted a simple verification experiment to gain some insights into the relative effectiveness of each sub-band and channel condition as a speaker classifying feature. The verification rate was calculated from the 2 PCD distributions (1,224 SS pairs and 373,320 DS pairs), the EER point used as the threshold, and the frequency regions with a higher speaker discriminability were sought. Each sub-band size and vowel combination was tested by altering sub-band ranges and frequency regions. The whole experiment was then repeated under 3 different channel conditions: microphone (Ch1), mobile phone (Ch3), and mismatch condition (Ch1-3).

Figure 2 presents a sample of the results taken from the tests using the sub-band range of 300Hz as Figure 1. Here, what immediately notable is the impact of channel mismatch. With every vowel and band range including the full-band, verification rate deteriorated considerably under mismatch conditions. Although this was expected, its significance is noteworthy.

The relationship between the two matching conditions, on the other hand, was somewhat unexpected. With all vowels but /o/, the Ch3 condition produced a better outcome than the Ch1 condition. Although it is counterintuitive, it has been reported that speech put through a mobile phone codec performs better in speaker classification than un-coded speech does [15], perhaps because the processing applied by mobile codec has the effect of reducing within-speaker variability. This does not agree with what we observed with F-ratios, however. Why the result for the vowel /o/ should be opposite to those of the other vowels remains unclear. However, with the /o/ vowels in Figure 1, F-ratio for Ch3 is notably worse than that for Ch1 at around 2200-3500Hz. We could speculate that the mobile transmission strongly affected the frequency region particularly relevant to speaker information for /o/ vowel (appears to be F3 region), and the positive effect of within-speaker variation reduction could not compensate this loss. Further investigation is needed, however.

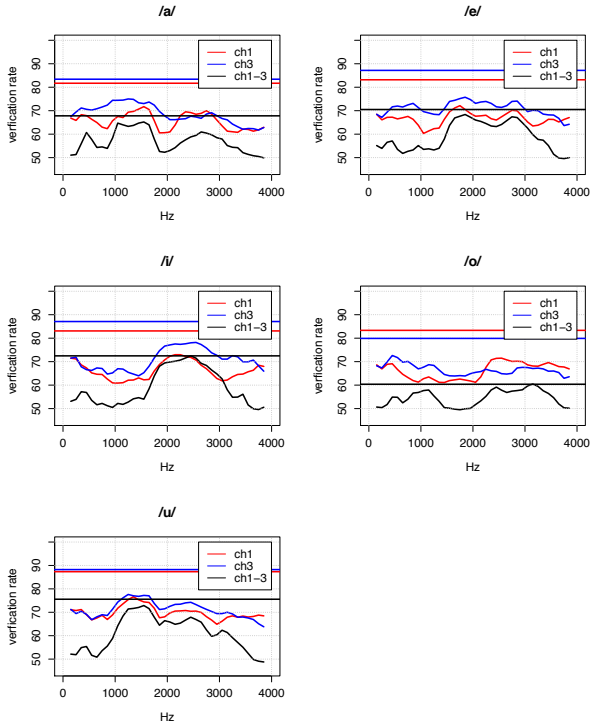


Figure 2: *Verification rates for varying frequency regions (sub-band 300Hz shifted by 100Hz step), horizontal lines indicating full-band results*

What is most notable with this experiment, however, is the relationship between the verification rates of the sub-band PCDs and the base-line full-band PCDs. The verification rate difference between two appears to be smaller under the mismatch condition in almost all the comparisons, suggesting that sub-band PCDs are more effective under mismatch conditions. This is a particularly exciting finding for FVC, as in most forensic cases, the two recordings to be compared

have been recorded on the different devices and transmitted through different channels. Telephone speech recordings are often compared to direct microphone recordings of police interviews. This channel mismatch has been long recognised as a hindrance to effective speaker classification. Various techniques for channel compensation have been proposed (e.g. [16-18]), but they all appear to require building a channel-characteristics model. However, since mobile transmission technology has an inherently highly variable signal processing path [15], the effectiveness of such approach may be limited. Further, crime-scene recordings are often very short. Thus the recording in question may not contain sufficient information to build a usable channel-characteristics model. Given all these constraints, it seems more practical to seek robust features against channel mismatch, rather than to attempt to compensate for this, at least in FVC casework contexts. The results from the current study seem to indicate that sub-band PCDs are promising features in this regard.

In observing the relationship between the full-band PCDs and the sub-band PCDs, we also found that, in particular combinations of the sub-band range and the frequency regions, the sub-band PCDs perform almost as well as the full-band PCDs, or occasionally better. This confirms the observation made in relation to F-ratios: the sub-band PCDs are likely to outperform the full-band PCDs in speaker classification, especially when multiple sub-bands are combined as partially independent sources of information.

5. Conclusion

This study explored the potential of the sub-band PCD as a speaker classification feature, using a large Japanese vowel dataset. Observations of F-ratios and verification rates revealed some promising characteristics of the sub-band PCDs. Firstly, the behaviour of sub-band PCDs is mostly predictable from our knowledge of articulatory and acoustic phonetics. This is all the more significant because of PCDs band-limited to formant ranges afford more direct articulatory interpretations than the typically-measured full-band cepstra. Secondly, sub-band PCDs seems more robust against channel mismatch than full-band PCDs. This is a welcome finding as most FVC casework involves speech data recorded under mismatch conditions.

The findings reported here warrant us to proceed to the next step: LR-based evaluation and FVC experiments based on this feature, which is presented as the second part of this study.

6. Acknowledgements

The work presented here was partly supported by JSPS KAKENHI Grant Number JP18H01671, JP25350488.

7. References

[1] G. S. Morrison, "Forensic voice comparison and the paradigm shift," *Science and Justice,* vol. 49, pp. 298-308, 2009.

[2] G. S. Morrison, "Tutorial on logistic-regression calibration and fusion:converting a score to a likelihood ratio," *Australian Journal of Forensic Sciences,* vol. 45, pp. 173-197, 2013/06/01 2012.

[3] G. S. Morrison, "Distinguishing between forensic science and forensic pseudoscience: Testing of validity and reliability, and approaches to forensic voice comparison," *Science & Justice,* vol. 54, pp. 245-256, 5// 2014.

[4] D. A. van Leeuwen and N. Brümmer, "An introduction to application - Independent evaluation of speaker recognition system," in *Speaker Classification*. vol. 1, C. Müller, Ed., ed Berlin: Springer, 2007, pp. 330--353.

[5] P. J. Rose, T. Osanai, and Y. Kinoshita, "Strength of forensic speaker identification evidence: Multispeaker formant and cepstrum-based segmental discrimination with a Bayesian likelihood ratio as threshold," in *The 9th Australian International Conference on Speech Science & Technology* Melbourne, 2002, pp. 303-308.

[6] 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, 2004, pp. 402-407.

[7] E. A. Alzqhoul, B. B. Nair, and B. J. Guillemin, "Comparison between Speech Parameters for Forensic Voice Comparison Using Mobile Phone Speech," in *The 15th Australasian International Conference on Speech Science & Technology*, Christchurch, 2014.

[8] C. Zhang, G. S. Morrison, F. Ochoa, and E. Enzinger, "Reliability of human-supervised formant-trajectory measurement for forensic voice comparison," *The Journal of the Acoustical Society of America,* vol. 133, pp. EL54-EL60, 2013.

[9] M. Duckworth, K. McDougall, G. de Jong, and L. Shockey, "Improving the consistency of formant measurement," *International Journal of Speech, Language & the Law,* vol. 18, pp. 35-51, 2011.

[10] G. K. Vallabha and B. Tuller, "Systematic errors in the formant analysis of steady-state vowels," *Speech Communication,* vol. 38, pp. 141-160, 9// 2002.

[11] F. Clermont and P. Mokhtari, "Frequency-band specification in cepstral distance computation," in *The 5th Australian International Conference on Speech Science & Technology* 1994, pp. 354-359.

[12] F. Clermont, Y. Kinoshita, and O. Takashi, "Sub-band cepstral variability within and between speakers under microphone and mobile conditions: A preliminary investigation," in *The 16th Australasian International Conference on Speech Science & Technology*, Sydney, 2016.

[13] H. Makinae, T. Osanai, T. Kamada, and M. Tanimoto, "Construction and preliminary analysis of a large-scale bone-conducted speech database," *IEICE technical report,* vol. Speech 107, pp. 97–102, 2007.

[14] Y. Kinoshita, "Testing Realistic Forensic Speaker Identification In Japanese: A Likelihood Ratio Based Approach Using Formants," PhD, Linguistics, The Australian National University, Canberra, 2001.

[15] E. A. Alzqhoul, B. B. Nair, and B. J. Guillemin, "Impact of dynamic rate coding aspects of mobile phone networks on forensic voice comparison," *Science & Justice,* vol. 55, pp. 363-374, 2015.

[16] A. A. Garcia and R. J. Mammone, "Channel-robust speaker identification using modified-mean cepstral mean normalization with frequency warping," in *Acoustics, Speech, and Signal Processing, 1999. Proceedings., 1999 IEEE International Conference on*, 1999, pp. 325-328.

[17] D. A. Reynolds, "Channel robust speaker verification via feature mapping," ed: I E E E, 2003, pp. II-53-6.

[18] A. Solomonoff, W. M. Campbell, and I. Boardman, "Advances in channel compensation for SVM speaker recognition," in *Acoustics, Speech, and Signal Processing, 2005. Proceedings.(ICASSP'05). IEEE International Conference on*, 2005, pp. I/629-I/632 Vol. 1.

Forensic voice comparison using sub-band cepstral distances as features: A first attempt with vowels from 306 Japanese speakers under channel mismatch conditions

Yuko Kinoshita[1], Takashi Osanai[2], Frantz Clermont[3]
[1] College of Arts and Social Science/Asia and the Pacific, The Australian National University
[2]National Research Institute of Police Science, Japan
[3]J.P. French Associates Forensic Lab., England
osanai@nrips.go.jp; yuko.kinoshita@anu.edu.au; dr.fclermont@gmail.com

Abstract

This study presents the latter part of an exploratory study of the potential of sub-band parametric cepstral distance (PCD) as an alternative forensic voice comparison (FVC) feature to formants and cepstral coefficients. Using 5 Japanese vowels produced by 306 male Japanese speakers, we conducted LR-based FVC experiments under a channel mismatch condition, with sub-bands selected in reference to the expected formant locations. Combining 3 sub-band PCDs from F1, F2, and F3 ranges, sub-band PCDs outperformed the full-band PCDs in speaker classification, demonstrating their promise as an automatically extractable, robust, and linguistically interpretable acoustic feature for FVC.

Index Terms: Sub-band cepstral distance, likelihood ratio, forensic voice comparison, channel mismatch, Japanese vowels

1. Introduction

Both speech and voice recognition systems are now part of our daily lives, and yet forensic voice comparison (FVC hereafter) is still no easy task. One of the reasons for this is the lack of control over the data. The speech samples in FVC, especially those from crime scenes, are often short and contain considerable background noise. The scarcity of data means insufficient data for modelling of speakers. Poor recording quality compromises the accuracy of acoustic feature extraction. As speakers are modelled based on those acoustic features, this also contributes to poor quality of the speaker models.

Also, the speech samples to be compared are most often recorded under very different circumstances. The speakers may be in very different emotional states, speak in different styles, and also be recorded on different devices via different transmission channels. These factors, which are unrelated to speaker characteristics, can amplify the acoustic differences between two samples, contributing to difficulties in producing strong likelihood ratios (LRs) in support of the same-speaker hypothesis, even where the speakers are indeed the same.

What the analyst can do to improve the situation is limited. We may be able to improve the elicitation and recording process of the known (or suspect) speaker, but crime scene recordings are largely out of our control.

Over the years, much research has been done on the impact of channel mismatch (e.g. [1-3]), and various techniques have been proposed to compensate for channel mismatch (e.g. [4-6]). However, such techniques all appear to require building a channel characteristics model. Crime scene recordings are often very short, so the recording in question may not contain sufficient information to build a reliable channel characteristics model. Also, mobile transmission characteristics change continuously, as the compression rate and methods change in response to network conditions [7, 8]. This makes the alternative approach, i.e. retrospectively 'matching' the conditions by putting a non-telephone recording (such as a police interview) through a mobile codec or a telephone network, less attractive. Further, various social network platforms now offer voice call options. It is thus increasingly unlikely for analysts to have access to full information on the processing applied to the speech sample in question.

These issues suggest that the most practical way forward in FVC is to search for features which are robust under forensically realistic conditions: less affected by external factors, and reliably measurable even with poor quality of recordings. This led us to the sub-band parametric cepstral distance (PCD), an approach initially proposed in [9]. PCD extracts the difference between two cepstral shapes within user-defined frequency boundaries. Its potential has been discussed in two studies: [10] examined within- and between-speaker variability of PCD, using landline telephone speech recordings from 297 Japanese speakers. Another study [11] made small scale observations on the F-ratio of sub-band PCDs using mobile and microphone recordings. The results from both studies were encouraging.

This motivated us to embark on the current project: an examination of the potential of sub-band PCD as an FVC feature using a large dataset. As the first step, we examined the behavior of sub-band PCDs in detail with respect to their F-ratios and verification rates in different sub-band ranges, using a database of Japanese vowels elicited from 306 speakers [12]. This database permits us to examine the forensically significant effect of channel mismatch, as it was recorded simultaneously via two channels: microphone and mobile phone transmission. The results of the initial experiments were promising; they suggested that sub-band PCD is relatable to articulatory gestures in similar ways to formants. This brings two advantages specific to forensic application: firstly, the results can be explained in court to non-experts in a relatively less abstract way; secondly any unusual results can be detected and reexamined in relation to articulatory and phonetic characteristics, more easily than full-band cepstra. They also found that speaker verification based on sub-band PCDs degrades less under a channel mismatch condition compared to that based on full-band PCDs, presumably because sub-band PCDs can exclude frequency ranges unrelated to speaker information.

This paper thus continues to examine the potential of sub-band PCDs as a speaker classification feature by selecting sub-

band ranges based on vowel formant frequencies, and conducting LR-based voice comparison experiments under a channel mismatch condition.

2. Data and procedures

2.1. Database, speakers, and speech materials

This study used the same data as [12]: 306 adult male speakers from the NRIPS database [13]. They are native speakers of Japanese, aged from 18 to 76 years. They had widely varied dialectal background, but dialectal variations appear not to affect vowel formants much in modern Japanese [14]. Thus the dialectal variation is unlikely to have contributed to greater between-speaker variability here. All speakers were recorded on two occasions, 2 to 3 months apart. They performed the same recording tasks twice at each recording session, and the whole process was recorded simultaneously through 2 channels: direct microphone (Ch1), and via a mobile phone network (Ch3). This study focuses on the cross-channel comparisons.

Read-out (C)V syllables were used as the speech samples: that is, Japanese 5 vowel phonemes, /a/, /e/, /i/, /o/ and /ɯ/, preceded by selected consonantal environment: ∅ (no consonant), /k/, /s/, /t/, /h/, /r/, /g/, /z/, /d/, /b/, and /p/. The phonemes /n/, /m/, /y/, and /w/ were excluded from analysis this time to facilitate reliable automatic segmentation. These are highly controlled elicitation, not spontaneous. However, [14] reports relatively small vowel reduction in running speech in Japanese. Therefore, we regard the current data as acceptable for this exploratory work.

Japanese *kana* syllabary writing system maintains the distinction between the pairs ぢ /di/ – じ /zi/ and づ /dɯ/ – ず /zɯ/, but they are phonetically identical, both realized as [dʑi] and [dzɯ]. Consequently, we have the vowel data in 10 different phonological contexts for /i/ and /ɯ/, and 11 for /a/, /e/, and /o/.

2.2. Segmentation and full-band LPCC extraction

The target syllables were automatically segmented into a preceding consonant and a vowel based on their power and F0. The sound files were down-sampled from 44.1 kHz to 8kHz, and full-band LPCCs were extracted from the selected vowel sections (order 14, Hamming window, window length 25ms, time-step 5ms). The LPCCs were averaged across the vowel duration, and further averaged across different phonological contexts for each vowel. As result, we obtained LPCCs for 5 vowels, 2 recording sessions, 2 repeats, and 2 recording channels for each speaker.

2.3. Parametric cepstral distance (PCD) calculation

The parametric cepstral distance (PCD) described in [9] affords selection of any sub-band range directly from full-band LPCCs. Its formulation is summarised below in Eq. (1), where $D^2\left(\overline{\mathbf{C}}_i, \overline{\mathbf{C}}_j, \omega_1, \omega_2\right)$ represents the Euclidean distance between any pair of full-band LPCCs $(\overline{\mathbf{C}}_i, \overline{\mathbf{C}}_j)$ for a given sub-band range. Note that the full-band LPCCs are index-weighted by the matrix **K** to emphasise spectral slope differences, and then weighted by the matrix $\mathbf{W}(\omega_1, \omega_2)$ to focus on any sub-band range selectable by its lower and upper limits ω_1 and ω_2. For $\omega_1 = 0$ and $\omega_2 = \pi$, Eq. (1) simply reduces to the familiar Euclidean distance between any pair of (index-weighted) full-band LPCCs.

$$D^2\left(\overline{\mathbf{C}}_i, \overline{\mathbf{C}}_j, \omega_1, \omega_2\right) = \left(\overline{\mathbf{C}}_i - \overline{\mathbf{C}}_j\right)^T \cdot \mathbf{K}^T \cdot \mathbf{W}(\omega_1, \omega_2) \cdot \mathbf{K} \cdot (\overline{\mathbf{C}}_i - \overline{\mathbf{C}}_j) \equiv \text{PCD between } \overline{\mathbf{C}}_i \text{ and } \overline{\mathbf{C}}_j \quad (1)$$

where:
$i, j \equiv$ speaker-session index
$\overline{\mathbf{C}}_i \equiv$ mean LPCC for i^{th} speaker across all tokens
$\overline{\mathbf{C}}_j \equiv$ mean LPCC for j^{th} speaker across all tokens

$$\mathbf{K} = \begin{bmatrix} \mathbf{1} & \mathbf{0} & \dots & \dots & \mathbf{0} \\ \mathbf{0} & \ddots & \ddots & & \vdots \\ \vdots & \ddots & \boldsymbol{k} & \ddots & \vdots \\ \vdots & & \ddots & \ddots & \mathbf{0} \\ \mathbf{0} & \dots & \dots & \mathbf{0} & \boldsymbol{M} \end{bmatrix} \equiv \text{index-weighting matrix}$$

$\mathbf{W}(\omega_1, \omega_2) \equiv$ band-selective matrix (see [9])
$\boldsymbol{M} \equiv$ LPCC order = 14
$\omega_1 \equiv$ lower limit of sub-band selected within $[0, \pi]$
$\omega_2 \equiv$ upper limit of sub-band selected within $[0, \pi]$

Sub-band PCD has the capacity to limit the analysis to the user-defined frequency regions, allowing us to exclude frequency regions that are unhelpful in assessing speaker identity. In the first part of this project, we found that the F-ratios tend to be higher in the frequency regions where we expect to find formants [12]. Thus, this time we select the sub-band frequency ranges referring to the formant measurements made in [15]. For each vowel, the mean ± 1 standard deviation of the first three formants were sought. The frequency ranges which contain the above values to the nearest 100Hz were defined as the target sub-band ranges for this study. These ranges are referred to as subF1, subF2 and subF3 hereafter.

Table 1. *Target sub-band ranges for each vowel (Hz).*

	subF1		**subF2**		**subF3**	
	from	to	from	to	from	to
/a/	600	800	1200	1600	2300	2800
/e/	300	600	1800	2200	2500	2900
/i/	200	400	1900	2400	2600	3100
/o/	300	600	1000	1300	2300	2700
/ ɯ /	200	400	1300	1800	2200	2700

2.4. Comparisons

With 306 speakers recorded 4 times (2 non-contemporaneous occasions, twice per sessions), we had 1224 patterns of same-speaker (SS) pairs and 3373320 patterns of different-speaker (DS) pairs. All comparisons were made in cross-channel conditions, i.e. between direct microphone recording (Ch1) and mobile phone network recording (Ch3).

2.5. Modelling and LR calculation

For LR calculation in linguistics-based FVC research, MVKD proposed by [16] has been a popular choice. It is, however, inappropriate to put PCDs through the MVKD formula, as a PCD is already a distance measure between two sets of information. The PCDs from the SS pairs and those from the DS pairs represent within- and between-speaker variations of the distance between two cepstra in the regions of the user-selected frequency ranges.

The relatively large data size in this study suggests that general population is reasonably well represented by the current data. However, examination of the PCD distributions revealed that the 1224 comparisons for SS were not sufficient to produce a smooth distribution, and direct derivation of LR

from it will result in some arbitrary fluctuations of LRs. The distributions need to be modelled.

To find an appropriate model, we tested the fit of four different distributions: normal, gamma, Weibull, and log normal, with the PCDs from different vowel and band combinations. Both SS and DS comparisons were evaluated for their fit to those 4 distributions based on Akaike's Information Criterion (AIC) and Bayesian Information Criterion (BIC). Although AIC and BIC measures fitting slightly differently [17], both measures selected an identical model as the best fit in all vowel and band range combinations in this study. Table 2 presents the counts of each model which produced the lowest AIC and BIC. SS and DS indicate the comparison types. The maximum score for each cell is 5, as we tested for all 5 vowels. Gamma clearly outperformed the rest.

Table 2: *Number of instances each model was selected as the best fit ()*

	norm		**gamma**		**weibull**		**lnorm**	
	SS	DS	SS	DS	SS	DS	SS	DS
Full	0	0	4	5	0	0	1	0
subF1	0	0	4	2	1	3	0	0
subF2	0	0	4	3	1	2	0	0
subF3	0	0	5	4	0	1	0	0
total	**0**	**0**	**17**	**14**	**2**	**6**	**1**	**0**

Based on this result, we fitted gamma distribution to the distributions of PCDs. Here, we added another type of PCD: sum of the PCDS from subF1, subF2 and subF3. This equates to the sum of area differences in three sub-band regions obtained from a pair of cepstra. Five vowels, 2 comparison types, and 3 sub-band ranges + full-band + summed PCD, resulted in 50 distributions. All were modelled with gamma distributions defined as below:

$$f(x) = \frac{1}{\Gamma(k)\theta^k} x^{k-1} e^{-x/\theta} \qquad (2)$$

We tested if lack of independence between testing and distribution modelling data has any effect by modelling the distributions with some speakers removed. We repeated 100 times removing a different set of 6 speakers each time, but no meaningful effect was found, as expected from the data size. Thus LRs were calculated by: 1) pooling PCDs separately for SS and DS comparisons and modelling their distributions, 2) deriving probabilities of the testing pairs to be belonging to the 2 different distributions, applying them to the models. The obtained LRs were then converted to Log_{10}LR (LLR) and calibrated. Cllr [18] was also calculated.

3. Results and Discussion

3.1. LLRs

In this section, we add another feature combination, summed LLR: the sum of 3 LLRs obtained from subF1, subF2, and subF3. Summing potentially correlated LLRs such as these risks introducing inaccuracy. However, the correlation among LLRs turned out to be very low, as seen in Table 3. The strongest correlation coefficient found was 0.188 (between subF1 and subF2 of /o/ vowel), indicating that correlations is unlikely to distort the results significantly.

Figure 1 presents the mean LLRs for each vowel and band range selection. It reveals that SS and DS comparisons are separated well at the theoretical threshold, LLR 0, across all vowels and band ranges. The vowel which produces strongest LRs — i.e. appearing at the furthest positions from 0 on both directions — is /ɯ/, closely followed by /i/. /a/ and /o/ appear to produce weaker LLRs. Comparing subF1, subF2 and subF3, we can see that subF1 is of limited use. The speaker information seems to be most richly carried in subF2.

Table 3: *Correlation between LLRs (Pearson's r)*

	SS comparisons			**DS comparisons**		
	F1-F2	*F1-F3*	*F2-F3*	*F1-F2*	*F1-F3*	*F2-F3*
a	0.072	-0.022	0.174	0.013	0.024	0.098
e	0.032	0.087	0.169	0.003	0.019	0.161
i	0.055	0.049	0.046	0.027	-0.003	0.131
o	0.188	0.128	0.045	0.045	0.109	-0.052
ɯ	0.102	0.121	0.102	0.040	0.053	0.055

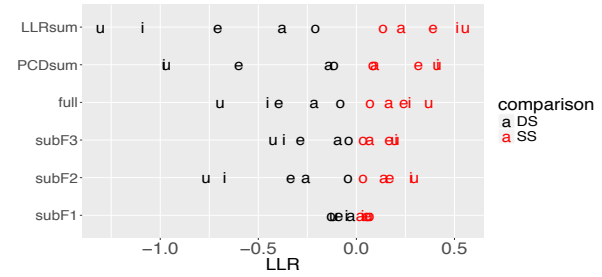


Figure 1: *Mean LLRs for each vowel and band range*

The results from the first part of this study [12] and the theoretical nature of the sub-band PCD predict sub-band PCDs (such as sum of PCD and sum of LLR) to outperform the full-band PCDs. Figure 1 shows that this is indeed to be the case; sum of PCD, and sum of LLR outperformed from full-band, confirming utility of band-selective analyses.

3.2. Verification rate and Cllr

Next, we observe the rates of successful speaker verification at threshold LLR 0. We focus our observation in this section on the comparison between full-band, sum of PCDs and sum of LLRs. With all vowels, the LLRs supported the correct hypothesis well above chance level, /i/ and /ɯ/ reaching over 80% for DS, which is a strong result for a single vowel. For all vowels but /a/, the sub-band based approach constantly outperformed full-band. Even for /a/, sum of LLR performed better than full-band. Here too the high vowels /i/ and /ɯ/ performed better than other vowels.

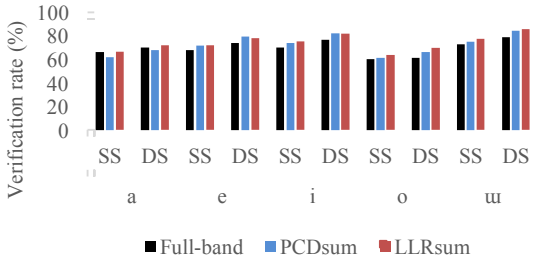


Figure 2: *Successful classification rate at LLR 0*

Cllr is the cost metric that evaluates the quality of the classification system [18]. We used half of the SS and DS comparisons for training the calibration, and the rest to

examine Cllr. Cllr can be decomposed to the verification cost (Cllr_min) and the calibration cost (Cllr_cal). For ease of interpretation, the decomposed components are presented in Figure 3. The results are presented separately for each vowel, /a/ to /ɯ/ from the bottom to the top. The categories "Full-band", "3_area", and "3_LLR" indicate full-band PCD, sum of PCD, and sum of 3 sub-band LLR.

Across all vowels, the scores for Cllr_cal (in dark blue) were very low, indicating that the system was already well calibrated, and the classification errors were largely caused by the PCD's discriminant capacity. The calibration results confirmed this; Cllrs did not improve with calibration, as seen in Figure 4, which presents the comparison of pre and post calibration Cllr, pooled across all band types in a violin plot.

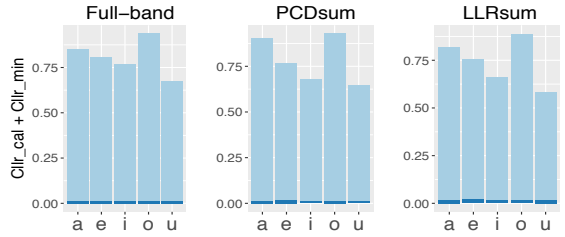


Figure 3: *Pre-calibration Cllr_min and Cllr_cal*

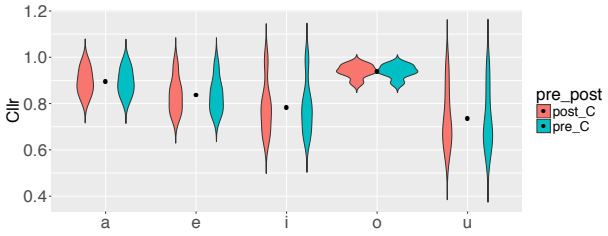


Figure 4: *Pre-and post-calibration difference in Cllr*

4. Conclusion

This paper further examined the behavior of sub-band PCDs. We selected 3 sub-bands for each vowel based on their known formant frequency ranges, and examined their performance under a channel mismatch condition. The results showed that individual sub-band PCDs were not as powerful as the full-band PCDs but once combined, they outperformed the full-band PCDs as predicted. Also LLRs produced from PCDs were found to be extremely well calibrated.

The examinations presented in [12] and here support our proposition of sub-band PCD being a potentially useful FVC feature. Most results were predictable from existing phonetic knowledge, suggesting sub-band PCD to be a feature that is automatically extractable and more readily interpretable – a desirable quality for evidence presentation in court.

As future tasks, performance comparison to the existing approaches is critical, especially to formant-based FVC. We also plan to do further work on optimal sub-band ranges, and the effect of sample data size and speech style, and different approaches to the LR calculation.

5. Acknowledgements

The work presented here was partly supported by JSPS KAKENHI Grant Number JP18H01671, JP25350488.

6. References

[1] H. J. Künzel, "Beware of the 'telephone effect': the influence of telephone transmission on the measurement of formant frequencies," *Forensic Linguistics* vol. 8, pp. 80-99, 2001.

[2] C. Byrne and P. Foulkes, "The 'mobile phone effect' on vowel formants," *International Journal of Speech Language and the Law,* vol. 11, pp. 83-102, 2004.

[3] 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, pp. 796-813, 7// 2013.

[4] A. A. Garcia and R. J. Mammone, "Channel-robust speaker identification using modified-mean cepstral mean normalization with frequency warping," in *Acoustics, Speech, and Signal Processing, 1999. Proceedings., 1999 IEEE International Conference on*, 1999, pp. 325-328.

[5] D. A. Reynolds, "Channel robust speaker verification via feature mapping," ed: I E E E, 2003, pp. II-53-6.

[6] A. Solomonoff, W. M. Campbell, and I. Boardman, "Advances in channel compensation for SVM speaker recognition," in *Acoustics, Speech, and Signal Processing, 2005. Proceedings.(ICASSP'05). IEEE International Conference on*, 2005, pp. I/629-I/632 Vol. 1.

[7] 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, 2008.

[8] E. A. Alzqhoul, B. B. Nair, and B. J. Guillemin, "Impact of dynamic rate coding aspects of mobile phone networks on forensic voice comparison," *Science & Justice,* vol. 55, pp. 363-374, 2015.

[9] F. Clermont and P. Mokhtari, "Frequency-band specification in cepstral distance computation," in *The 5th Australian International Conference on Speech Science & Technology* 1994, pp. 354-359.

[10] M. Khodai-Joopari, F. Clermont, and M. Barlow, "Speaker variability on a continuum of spectral sub-bands from 297-speakers' non-contemporaneous cepstra of Japanese vowels," in *The 10th Australian International Conference on Speech Science and Technology*, Sydney, 2004, pp. 504-509.

[11] F. Clermont, Y. Kinoshita, and O. Takashi, "Sub-band cepstral variability within and between speakers under microphone and mobile conditions: A preliminary investigation," in *The 16th Australasian International Conference on Speech Science & Technology*, Sydney, 2016.

[12] T. Osanai, Y. Kinoshita, and F. Clermont, "Exploring sub-band cepstral distances for more robust speaker classification," presented at the 17th Speech Science and Technology Conference (SST2018), Sydney, 2018.

[13] H. Makinae, T. Osanai, T. Kamada, and M. Tanimoto, "Construction and preliminary analysis of a large-scale bone-conducted speech database," *IEICE technical report,* vol. Speech 107, pp. 97–102, 2007.

[14] 奥田浩三, "発話スタイルの変動に頑健な音響モデル構築法に関する研究," 大阪市立大学, 2005.

[15] Y. Kinoshita, "Testing Realistic Forensic Speaker Identification In Japanese: A Likelihood Ratio Based Approach Using Formants," PhD, Linguistics, The Australian National University, Canberra, 2001.

[16] C. Aitken, G.G. and D. Lucy, "Evaluation of trace evidence in the form of multivariate data," *Applied Statistics,* vol. 53, pp. 109--122, 2004.

[17] S. I. Vrieze, "Model selection and psychological theory: A discussion of the differences between the Akaike information criterion (AIC) and the Bayesian information criterion (BIC)," *Psychological Methods,* vol. 17, pp. 228-243, 2012.

[18] D. A. van Leeuwen and N. Brümmer, "An introduction to application - Independent evaluation of speaker recognition system," in *Speaker Classification*. vol. 1, C. Müller, Ed., ed Berlin: Springer, 2007, pp. 330--353.

Independent Modelling of Long and Short Term Speech Information for Replay Detection

Gajan Suthokumar[1,2], Kaavya Sriskandaraja[1], Vidhyasaharan Sethu[1], Chamith Wijenayake[1], Eliathamby Ambikairajah[1,2]
[1]School of Electrical Engineering and Telecommunications, UNSW Australia
[2]DATA61, CSIRO, Sydney, Australia
g.suthokumar@unsw.edu.au

Abstract

Cepstral normalisation is widely employed in replay detection systems. However, incorporating some information that is lost during normalisation may be useful. Additionally, anti-spoofing systems may further benefit from not treating all speech frames identically. In this paper, we separate speech information based on two different criteria and model them independently. Three novel approaches are proposed based on (1) long and short term information; (2) high and low energy frames; and (3) a combination of the two. Experiments were conducted on the ASVSpoof2017 (V2.0) corpus and the best results correspond to an EER of 8.67% with a relative improvement of 29%.

Index Terms: speaker verification, spoofing detection, replay

1. Introduction

Significant improvements have been made in automatic speaker verification (ASV) in recent decades; however, concerns about security vulnerabilities continue to form a barrier to their widespread use. Speaker verification uses voice biometrics to verify the claimed speaker from a given speech utterance [1]. ASV systems are vulnerable to a diverse range of spoofing attacks, including speech synthesis (SS), voice conversion (VC), impersonation and replay [1], all of which have been shown to heavily degrade the robustness of ASV. Replay attacks, the playback of pre-recorded genuine speech, are arguably the most common ASV spoofing technique since they do not require attackers to have any special speech technology knowledge and can be mounted with relative ease using common consumer devices.

Developing anti-spoofing techniques to effectively mitigate the replay attacks generally aims to exploit one of the several factors: the fact that replayed speech would be a copy of a previous speech utterance [2]; differences in the transmission channel; or differences in the spectral properties of replayed speech. In literature, transmission channel differences are identified in the form of pop-noise [3], acoustic channel artefacts [4], and the detection of far-field recording [5]. Most of other techniques are based on the short term spectral cue differences. These include rectangular filter cepstral coefficients (RFCCs) [6] , spectral centroid magnitude coefficients (SCMCs) [6] , constant-Q cepstral coefficients (CQCCs) [7], scattering cepstral coefficients (SCCs) [8] and inverse Mel frequency cepstral coefficients (IMFCCs) [9]. In addition to the spectral based features, phase [10] and voice source features [10] have also been investigated. Spectral cues captured by the short term features derived from a linear frequency scale have dominated over warped frequency scales [6]. Different variants of neural network (NN) based systems [11] have also been investigated. Gaussian mixture models (GMMs) remain superior to other classifiers [6], [11]. Apart from these individual features and classifiers, the literature also indicates that the score fusion of different types of features and systems can perform better in replay detection.

Cepstral normalisation shown to be highly beneficial in replay detection in the form of cepstral mean normalisation (CMN) [6] and cepstral mean and variance normalisation (CMVN) [12], which normalise the temporal cepstral statistics (mean and variance) of short term (ST) features across each dimension. Even though, the cepstral normalisation of features in replay detection may at first seem counter-intuitive, the study of [12] argued that this may help to align both genuine and replayed speech distributions on to a similar scale. However, cepstral normalisation might also remove the information that could be useful for replay detection. Because, normalisation techniques are used in many other speech applications to normalise nuisance channels [13][14], which arise due to heterogeneous conditions, e.g. recording device and environment. Also, long term (LT) spectral statistics have been shown to be effective in spoofing detection [15]. In addition to that, temporal features based on amplitude modulation have demonstrated the significance of the long term temporal dynamics in our previous work [16].

Apart from that, standard replay detection systems model all the frames identically, however voiced and unvoiced regions could mask channel information differently and a speaker's voice might mask the channel information in voiced regions, so if unvoiced portions are focused on, the channel information may become more pronounced [17]. This further suggests that voiced and unvoiced frames might not contain identically emphasized discriminative information for replay detection.

In this paper, we propose systems based on the following two hypotheses: (a) Cepstral statistics of short term features that are removed during normalisation may contain discriminative information for replay detection; and (b) voiced and unvoiced region artefacts might not be emphasized in the same manner in the presence of replay channels. Specifically, we propose separating speech into regions with non-overlapping and complimentary information and modelling the differences between replayed and genuine speech in these regions independently and that are then fused at the score level. It is noted that the state-of-the-art replay detection systems model the cepstral normalised (short term) features only.

2. Proposed System Architectures

We proposed three architectures to individually model the distribution of (1) cepstral normalised features (short term) and cepstral statistics (long term) as shown in Figure 2; (2) high energy (HE) and low energy (LE) frames as shown in Figure 4, to form two spoofing detection systems that are then fused at the score level. The third proposed hybrid architecture

is a combination of the independent parallel paths of high energy, low energy and cepstral statistics of the short term feature information.

2.1. Short and Long Term Based Separation

In order to determine the effectiveness of the cepstral mean and variance in replay detection, a t-SNE representation of cepstral mean and cepstral variance of RFCCs from genuine (green) and spoofed speech (orange) is shown in Figure 1, where the feature spaces within genuine and spoof classes of the entire training set is compared. It can be seen that the feature space of the cepstral mean of RFCC feature depicts good discriminability, while the cepstral variance of RFCC feature space shows less. It is arguable that the incorporation of either cepstral mean or cepstral variance (long term) of the RFCC feature independently with the cepstral normalised RFCC features (short term) as depicted in Figure 2 could be helpful for improved replay detection.

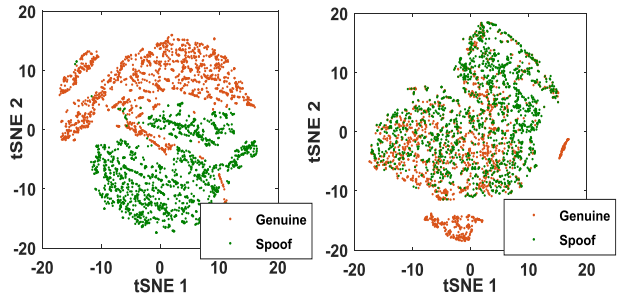


Figure 1: *t-SNE plot for genuine (orange) and spoof (green) RFCCs for Cepstral Mean (left) and Cepstral Variance (right) on the entire training set.*

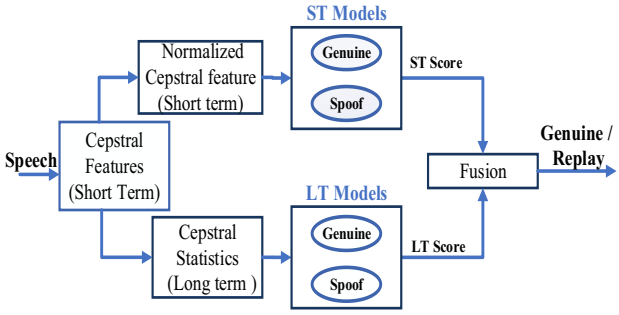


Figure 2: *System architecture based on short-term (ST) and long-term (LT) separation*

2.2. High Energy (HE) and Low Energy (LE) Based Separation

Speech frames are initially categorised as either high energy (HE) or low energy (LE) frames using a voice activity detector (VAD). Here we expect the LE frames to contain unvoiced speech. In order to determine effectiveness of separating high and low energy frames, we show the t-SNE plots of the RFCCs from genuine and spoofed speech in terms of HE and LE frames in Figure 3 and compare the feature spaces of genuine and spoofed speech. It can be seen that the features for genuine HE and LE frames occupy different spaces, and a similar pattern is observed for spoofed speech as well. Thus, it is also arguable that separate modelling of HE and LE frames could be helpful for improved replay detection, with an architecture as depicted in Figure 4. Similar technique has proved to be effective in synthetic speech detection in our previous work [18] since the voiced speech and unvoiced speech are not synthesized in same manner in speech synthesis and voice conversion algorithms.

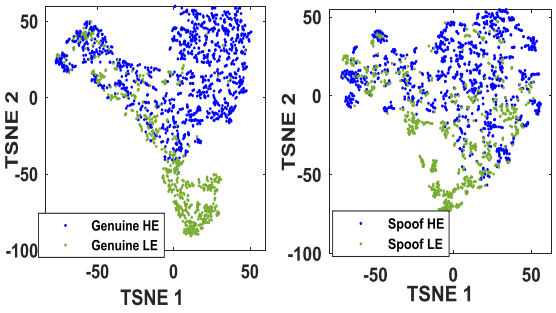


Figure 3: *t-SNE plots of the RFCC static features from a subset of the training set for genuine (left) and spoofed speech (right), in terms of high energy (HE) (blue) and low energy (LE) (green) frames.*

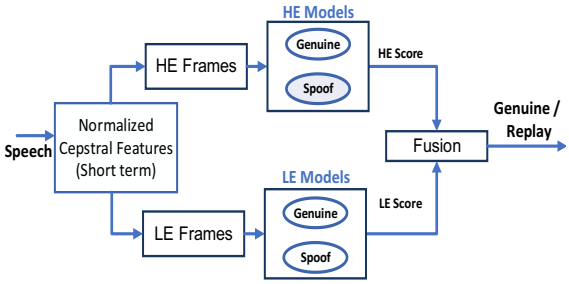


Figure 4: *System architecture based on high energy (HE) and low energy (LE) separation.*

2.3. Hybrid Architecture

A depiction of the proposed hybrid architecture is shown in Figure 5, to combine the individual gains of both of the proposed architectures.

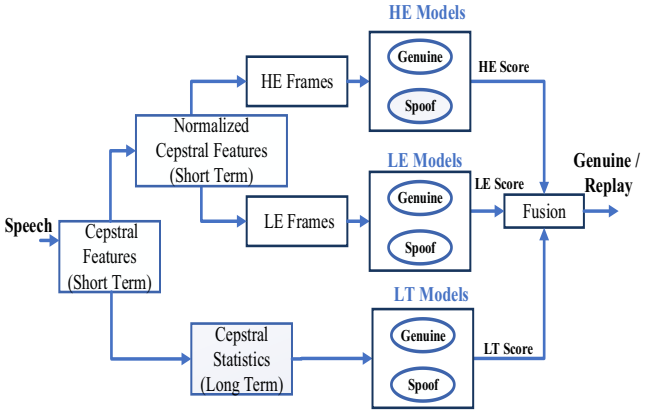


Figure 5: *Hybrid system.*

3. Experimental Setup

3.1. Database

The ASVspoof 2017 database consists of genuine recordings and their replayed versions as spoofed speech. Genuine utterances are sourced from the RedDots corpus. Spoofed utterances are created through recording and replaying genuine utterances using a variety of recording devices, playback devices and acoustic environments. This database consists of three non-overlapping subsets for training, development and evaluation. The ASVspoof 2017 Version 2.0 (V2.0) database [12] and the CQCC baseline are released earlier this year by the challenge organisers, is an updated version of the original ASVSpoof 2017 Challenge database Version 1.0 (V1.0) [7], correcting several data anomalies found in the original. As such, previously reported results using the V1.0 database are not directly comparable with the V2.0 results. In addition to this, the meta data of recording and playback devices, and the environmental conditions of the evaluation set is also released in V2.0. All reported experiments in this paper are conducted on the V2.0 database.

3.2. Voice Activity Detection

Vector quantization based VAD (VQVAD) [19] was employed in preference to other VAD's because of its unsupervised nature as well as it does not depend on a pre-determined threshold. Also, it shows better performance on short duration utterances which is suitable for ASVspoof 2017 V2.0 corpus [19]. The VQVAD is tuned for the ASV microphone environment with default parameters. HE frames are expected to include voiced frames, while LE frames are expected to include unvoiced frames and silence. The ASVSpoof 2017 V2.0 corpus consists of 54% HE frames and 46% LE Frames.

3.3. Front-End

RFCCs [6] and SCMCs [6] were used as the front-end for our experiments, as they are the state-of-the-art short term features for replay attack detection. They are extracted with a frame size of 20ms with 50% overlap. 40 dimensional static and dynamic features (i.e velocity and acceleration) are utilized. CMN is carried out for all our short term features unless otherwise specified.

3.4. Classifier and Score level Fusion

A 2-class GMM back-end was employed for genuine and spoofed speech detection. The GMMs were trained using the expectation maximization (EM) criterion, for genuine and spoofed speech with random initialization. The proportion of the amount of HE and LE frames in development set is considered in the selection of number of GMM mixtures. We investigated a range of numbers of GMM mixtures, eventually employed 512 mixtures for the baseline (i.e. no separation) and chose 256 each for the HE and LE models. A small number of mixtures to model the cepstral statistics (LT) is sufficient as it is an utterance level feature and 4 mixtures are found to be optimal. A linear regression based score level fusion from the Bosaris toolkit [20] is employed in order to combine the independent classifier scores, since the features associated with each models are highly complementary.

4. Results and Discussion

4.1. Long term Cepstral Statistics Features

Table 1 shows the development set performance for the long term cepstral statistics features (i.e. mean and variance) of static (S) and combined static and dynamic (i.e velocity and acceleration) (SD) for RFCCs and SCMCs. Firstly, the systems using cepstral mean features (i.e. LT_M (S) & LT_M (SD)) consistently performed better than those using cepstral variance features (i.e. LT_V (S) & LT_V (SD)). Secondly the cepstral means systems, LT_M (S) performs better than LT_M (SD), in contrast to the variance systems, while LT_V (SD) performs better than LT_V (S). The success of the cepstral statistics features is reasonable as the temporal dynamics of the replayed speech tend to be affected, presumably due to the heterogenous replay channels and environments.

Table 1. *Development results in terms of % EER for individual long term cepstral mean (LT_M) & variance (LT_V) systems, for RFCC and SCMC static (S) features, and combined static & dynamic (SD) features.*

System	RFCC	SCMC
LT_M (S)	**13.65**	**13.64**
LT_V (S)	22.53	23.04
LT_M (SD)	**18.34**	**16.84**
LT_V (SD)	20.68	21.20

4.2. Comparative Performance

Table 2 shows the development set performance for different combinations of the proposed systems and the baseline systems for RFCCs and SCMCs. It is noted that, our baseline front-ends are better than the improved CQCC system [12] released with ASVSpoof 2017 V2.0. The proposed HE+LE, ST+LT_M, and hybrid systems outperform the baseline, which models only the normalised short term features. Again, the incorporation of the static LT_M (S) features seems superior to the LT_M (SD).

Table 3 shows the evaluation set performance for different combinations of the proposed best systems identified on the development set, as well as the baseline systems for RFCC and SCMC features. All of the proposed systems are superior to the baseline system, which does not independently model neither the separated speech information nor the cepstral statistics.

Table 2. *Development set results in terms of %EER for RFCC and SCMC features for the proposed systems.*

Architecture	System	RFCC	SCMC
Baseline	ST [6]	7.76	8.66
Proposed 1	ST+ LT_M(S)	6.50	7.35
	ST+ LT_M(SD)	6.68	7.68
Proposed 2	HE+LE	7.15	8.41
Proposed 3	HE+LE+ LT_M(S)	**6.12**	**6.99**
	HE+LE+LT_M(SD)	6.38	7.27

Table 4 compares the evaluation set performances of our best proposed system (HE+LE+LT_M(S)) with the previously

reported best results [12], for different threat conditions as defined in [12]. The meta data of ASVSpoof 2017 V2.0 defines the different replay threat conditions for recording device, playback device and acoustic environments. It can be noticed that the proposed system is superior under all threat conditions.

Table 3. *Evaluation set results in terms of %EER for RFCC and SCMC features for the proposed systems.*

Architecture	System	RFCC	SCMC
Baseline	ST [6]	11.22	12.23
Long Term	LT_M(S)	16.96	17.05
	LT_M(SD)	16.94	15.46
Proposed 1	ST+ LT_M(S)	**10.28**	**10.09**
	ST+ LT_M(SD)	11.20	11.65
Proposed 2	HE+LE	10.42	11.01
Proposed 3	HE+LE+ LT_M(S)	**9.03**	**8.67**
	HE+LE+LT_M(SD)	9.88	9.82

Table 4. *Proposed best systems results in terms of % EER of RFCC/SCMC features for different threat conditions (low, medium and high) as defined in* [12] *in terms of % EER. (The best results previously reported on version 2.0* [12] *are given within parentheses).*

Conditions	Low	Medium	High
Environment	**8.44/8.28** (16.68)	**8.13/7.29** (18.73)	**14.52/13.13** (21.86)
Playback Device	**8.24/8.58** (16.64)	**7.04/6.6** (16.44)	**10.65/10.28** (18.37)
Recording Device	**7.77/7.27** (10.80)	**8.24/8.01** (15.69)	**9.99/9.73** (17.77)

5. Conclusions

A novel framework is proposed in this paper to independently model speech information separated based on two different criteria, prior to score fusion. This approach places a greater emphasis on relevant speech information compared to the standard approach. This is beneficial since this information encompasses complementary discriminative ability for replay detection, which is not well emphasized in the standard approach whereby all genuine and spoofed speech information is described by one GMM each. The incorporation of the long term cepstral statistics of the short term features that are discarded during cepstral normalization is proved to be beneficial for replay detection. The approach of independently modelling the high energy and low energy frames and cepstral statistics of the short term feature was found to be superior to the standard approach. The proposed framework has been validated on the ASVSpoof 2017 V2.0 corpus and the results consistently showed that the proposed approach is superior to the standard approach with a 29% relative improvement.

6. References

[1] Z. Wu, N. Evans, T. Kinnunen, J. Yamagishi, F. Alegre, and H. Li, "Spoofing and countermeasures for speaker verification: A survey," *Speech Commun.*, vol. 66, pp. 130–153, Feb. 2015.

[2] Z. Wu, S. Gao, E. S. Cling, and H. Li, "A study on replay attack and anti-spoofing for text-dependent speaker verification," *2014 Asia-Pacific Signal Inf. Process. Assoc. Annu. Summit Conf. APSIPA 2014*, pp. 92–96, 2014.

[3] S. Shiota, F. Villavicencio, J. Yamagishi, N. Ono, I. Echizen, and T. Matsui, "Voice Liveness Detection for Speaker Verification Based on a Tandem Single / Double-Channel Pop Noise Detector," pp. 259–263, 2016.

[4] Z.-F. Wang, G. Wei, and Q.-H. He, "Channel pattern noise based playback attack detection algorithm for speaker recognition," in *2011 International Conference on Machine Learning and Cybernetics*, 2011, pp. 1708–1713.

[5] J. Villalba and E. Lleida, "Preventing replay attacks on speaker verification systems," in *2011 Carnahan Conference on Security Technology*, 2011, pp. 1–8.

[6] R. Font, J. M. Espín, and M. J. Cano, "Experimental Analysis of Features for Replay Attack Detection — Results on the ASVspoof 2017 Challenge," in *Interspeech*, 2017, pp. 7–11.

[7] T. Kinnunen *et al.*, "ASVspoof 2017: Automatic Speaker Verification Spoofing and Countermeasures Challenge Evaluation Plan," vol. 0, no. 1, pp. 1–5, 2016.

[8] K. Sriskandaraja, G. Suthokumar, V. Sethu, and E. Ambikairajah, "Investigating the use of scattering coefficients for replay attack detection," in *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2017, pp. 1195–1198.

[9] M. Witkowski, S. Kacprzak, P. Zelasko, K. Kowalczyk, and J. Gałka, "Audio Replay Attack Detection Using High-Frequency Features," in *Interspeech*, 2017, pp. 27–31.

[10] S. Jelil, R. K. Das, S. R. M. Prasanna, and R. Sinha, "Spoof Detection Using Source , Instantaneous Frequency and Cepstral Features," in *Interspeech*, 2017, pp. 22–26.

[11] G. Lavrentyeva, S. Novoselov, E. Malykh, A. Kozlov, O. Kudashev, and V. Shchemelinin, "Audio replay attack detection with deep learning frameworks," in *Interspeech*, 2017, pp. 82–86.

[12] M. Todisco, N. Evans, T. Kinnunen, K. A. Lee, and J. Yamagishi, "ASVspoof 2017 Version 2.0: meta-data analysis and baseline enhancements," in *Odyssey*, 2018.

[13] O. Strand and A. Egeberg, "Cepstral mean and variance normalization in the model domain," *COST278 ISCA Tutor. Res. Work. Robustness Issues Conversational Interact.*, pp. 2–5, 2004.

[14] O. Viikki and K. Laurila, "Cepstral domain segmental feature vector normalization for noise robust speech recognition," *Speech Commun.*, vol. 25, pp. 133–147, 1998.

[15] H. Muckenhirn, P. Korshunov, M. Magimai-Doss, and S. Marcel, "Long-Term Spectral Statistics for Voice Presentation Attack Detection," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 25, no. 11, pp. 2098–2111, 2017.

[16] G. Suthokumar, V. Sethu, C. Wijenayake, and E. Ambikairajah, "Modulation Dynamic Features for the Detection of Replay Attacks," in *Interspeech*, 2018, pp. 691–695.

[17] Z. H. Lim, X. Tian, W. Rao, and E. S. Chng, "An investigation of spectral feature partitioning for replay attacks detection," in *2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2017, pp. 1570–1573.

[18] G. Suthokumar, K. Sriskandaraja, V. Sethu, C. Wijenayake, and E. Ambikairajah, "Independent Modelling of High and Low Energy Speech Frames for Spoofing Detection," in *Interspeech*, 2017, pp. 2606–2610.

[19] T. Kinnunen and P. Rajan, "A practical, self-adaptive voice activity detector for speaker verification with noisy telephone and microphone data," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2013, pp. 7229–7233.

[20] N. Brümmer and E. de Villiers, "The BOSARIS Toolkit: Theory, Algorithms and Code for Surviving the New DCF," Apr. 2013.

Supervised Variational Relevance Learning (SUVREL) applied to voice comparison

Eduardo R. Silva, Manfredo H. Tabacniks

Department of Applied Physics, University of Sao Paulo, BR

edurs@if.usp.br, tabacniks@if.usp.br

Abstract

An introductory survey of the application of the Supervised Variational Relevance Learning (SUVREL), as a preprocessing tool, to voice comparison and classification was made in conjunction with the Principal Component Analysis (PCA) classifier. SUVREL's algorithm produces an explicitly obtained metric tensor with low computational cost which can be possibly suitable to perform speaker identification over low processing devices, for instance, cell phones. The results demonstrate a considerable increase in the total variance explained by PCA's first three components obtained before and after preprocessing with SUVREL, on the original quality data and after addition of white noise: 73.17% to 99.87% and 79.15% to 99.96% for data from 100 male subjects and 71.09% to 99.84% and 76.22% to 99.93% for data from 100 female subjects. Using an ellipsoidal representation of each subject in the PCA's space of scores, the volume intersection was evaluated as an indicator of subject's voice similarities. Percentual decrease in the overall volume intersection after preprocessing ranged from 6.65 ± 0.05 % to 20.44 ± 0.05 %.

Index Terms: SUVREL, relevance learning, distance learning, speech signal processing.

1. Introduction

The performance of many algorithms for classification depends fundamentally on them being given a good metric over the input space. Suport Vector Machines (SMV), K-means and several forms of supervised Malahanobis distance learning were developed to achieve this objective [1]. The present work consists of an introductory survey of the application of the method Supervised Variational Relevance Learning (SUVREL) to voice comparison and classification of a database of voice recordings.

2. SUVREL: Supervised Variational Relevance Learning

The *Supervised Variational Relevance Learning* (SUVREL) consists in obtaining a metric tensor which allows pattern identification based on similarities [2]. In literature, there are several models and functions which penalize features that increase intraclass distances and favor small interclass distances (e.g. [1] and [3]). SUVREL method presents a cost function which also penalize features that increase intraclass distances and favor small interclass distances but with the computational advantage that it can be obtained analytically. SUVREL aims to be a preprocessing tool that increase data classification efficiency. SUVREL is a supervised method since it requires previous knowledge in which class each element is contained.

Consider an experiment i and its correspondent data vector $\vec{x_i} = (x_{i1}, \cdots, x_{iF})$ in the space of features $\mathbb{R}^F$, and each feature $\mu = 1, \cdots, F$ representing a dimension. Learning and validation sets $\mathcal{L}$ and $\mathcal{V}$ are organized in sets of pairs $\{\vec{x_i}, c_m\}$ – or concisely $\vec{x_i^{c_m}}$ – where c_m are class labels. The set $\mathcal{C} = \{c_1, c_2, \cdots, c_m\}$ refers to the m classes of the experiment, with $m = \{1, 2, \cdots, k\}$, and the total number of elements n in a set is given by:

$$n = \sum_{m=1}^{k} n_{c_m} \tag{1}$$

where n_{c_m} is the total number of elements in the m-th class.

One way to define a geometry in the space of features is to impose a structure through a metric tensor $g_{\mu\nu}$. Let the distance between two elements i e j be defined by:

$$d_{ij}^2 = \sum_{\mu\nu} g_{\mu\nu} \Delta x_{ij\mu} \Delta x_{ij\nu} \tag{2}$$

where $\Delta x_{ij\nu} = x_{i\nu} - x_{j\nu}$ is the difference between patterns i e j for the ν-th feature.

SUVREL model introduces the cost function:

$$E(\{g_{\mu\nu}\}; \mathcal{L}) = \sum_{a \in \mathcal{C}} \langle d_{ij}^2 \rangle_{a,a} - \gamma \sum_{<a \neq b> \in \mathcal{C}} \langle d_{ij}^2 \rangle_{a,b} \tag{3}$$

where a and b are class labels indices expliciting intra and interclasses operations and γ is a parameter to control the weigh given to the relation between inter/intraclasses distances.

The average distance d_{ij}^2 is defined as:

$$\langle d_{ij}^2 \rangle_{a,b} = \frac{1}{n_{ab}} \sum_{i \in a, j \in b} d_{ij}^2 \tag{4}$$

where $n_{ab} = n_a n_b$ and $\sum_{i \in \mathcal{C}}$ is the sum over all elements pertaining to class a.

Cost function 3 can be rewritten as:

$$E = \sum_{\mu\nu} g_{\mu\nu} \epsilon_{\mu\nu} \tag{5}$$

where:

$$\epsilon_{\mu\nu} = e_{\mu\nu}^{in} - \gamma e_{\mu\nu}^{out} \tag{6}$$

with terms:

$$e_{\mu\nu}^{in} = 2 \sum_{a \in \mathcal{C}} cov(\vec{x_\mu^a}; \vec{x_\nu^a}) \tag{7}$$

and:

$$e_{\mu\nu}^{out} = (k-1) \sum_{a \in \mathcal{C}} cov(\vec{x_\mu^a}; \vec{x_\nu^a}) + k^2 cov(\mathbf{m}_\mu; \mathbf{m}_\nu) \tag{8}$$

where it is defined:

$$\mathbf{m}_\mu = \{m_{a\mu}\} = \left\{ \frac{1}{n_a} \sum_{i \in a} x_{i\mu} \right\} \tag{9}$$

and :

$$cov(\mathbf{m}_\mu;\mathbf{m}_\nu) = \frac{1}{k}\sum_{a\in\mathcal{C}} m_{a\mu}m_{a\nu} - \left(\frac{1}{k}\sum_{a\in\mathcal{C}} m_{a\mu}\right)\left(\frac{1}{k}\sum_{a\in\mathcal{C}} m_{a\nu}\right) \tag{10}$$

Thus:

$$\epsilon_{\mu\nu} = [2-(k-1)\gamma]\sum_{a\in\mathcal{C}} cov(\vec{x_\mu^a};\vec{x_\nu^a}) - \gamma k^2 cov(\mathbf{m}_\mu;\mathbf{m}_\nu) \tag{11}$$

which just depends on data vectors and the provided information about to which experiment each class a refers.

Relevance is obtained by applying the method of Lagrange multipliers to minimize (5) with respect to the metric tensor $g_{\mu\nu}$, subject to the scale constraint $\sum_{\mu\nu} g_{\mu\nu}^2 = 1$, with a fixed γ. Solving:

$$\frac{\delta}{\delta g_{\mu\nu}}\left[E+\theta\left(\sum_{\mu\nu} g_{\mu\nu}^2 - 1\right)\right] = 0 \tag{12}$$

it is obtained:

$$g_{\mu\nu} = \frac{-\epsilon_{\mu\nu}}{\sum_{\mu'\nu'} \epsilon_{\mu'\nu'}^2} \tag{13}$$

The resolution analytically obtained reduces considerably computational time in comparison to numerical methods. In [2], it is proved in details that the necessary condition to get a positive definite tensor is that γ be greater than a γ^* defined as:

$$\gamma^* < \frac{2}{(k-1)} \tag{14}$$

Since SUVREL analytically computes a metric tensor, therefore with lower computational cost in comparison with numerically obtained metrics, the study of SUVREL's effects in conjunction with a classifier such as Principal Component Analysis, Multidimensional Scaling (MDS) etc. for voice comparison may be promising, making its application possibly suitable to perform speaker identification over low processing devices, for instance, cell phones.

3. Methods

This section describes the database employed, the data extraction and its analysis.

In this experiment, it was used a database of voice recordings of more than 500 Australian English speakers [4] for forensic-voice-comparison research which collection protocol is described in details in [5]. This database consists of three blocks of recordings where speakers were asked to perform three speaking tasks: informal telephone conversation, information exchange over the telephone and pseudo-police-style interview. Results here obtained were made over the first block of recordings, with both males and females subjects. Each record was used in its original sample frequency 44 100 Hz. There are comparative results for the original high quality recordings and after white Gaussian noise addition (SNR = 25 dB)[1].

Using a script developed in Matlab R2016a®, 44 Mel-frequency cepstral coefficient (MFCC) acoustic features were extracted for each labelled individual for samples of about 30 seconds divided into frames of 25 milliseconds with an overlap of 10 milliseconds, from which 2nd to 43rd MFCC were retained. The triangular Mel-filter bank used here is defined as in [6, pp. 351].

[1] Signal to Noise Ratio.

The data formed with each set of MFCCs, already class labelled accordingly to each subject, was preprocessed with SUVREL algorithm. The complete metric tensor $g_{\mu\nu}$ (eq. 13) – rather than only its principal diagonal – calculated for a fixed $\gamma = 2$ (obbeying eq. 14) was used. Both original and preprocessed data were classified using Principal Component Analysis (PCA) [7]. The percentage of the total variance explained by each PCA component were then compared. PCA's score plots of the first three components were made to show the distribution of the MFCCs rescaled in PCA's space.

Since a direct plot of PCA's scores would just lead to a visually incomprehensible cloud of dots, and to achieve a better understanding of the results, each subject's "subcloud" were represented by an ellipsoid centered at the subcloud's mean value and stretched proportionally to its standard deviation, always calculated with respect to each of the first three PCA components. This way, each individual occupies a volume portion of the PCA's score space and similarities between each other subjects can be numerically estimated by ellipsoid's volumes intersections. Volume intersection estimations were calculated with Monte Carlo's algorithm, with error below 0.05%. To illustrate this representation, the plot presented in fig. 3 was drawn with just a fourth of the standard deviations to accomplish a better visualization – although all calculations were made using total standard deviations.

4. Results

MFCC data was obtained from 30 seconds sample utterances (as described in sec. 3) from each one of 100 male subjects. After addition of white Gaussian noise at SNR = 25 dB to these same utterances, MFCCs were obtained again. These data sets were preprocessed with SUVREL. All data were classified via Principal Component Analysis. Table 1 presents the percentage of the total variance explained by the first three PCA components. A considerable improvement after preprocessing with SUVREL and a concentration of almost all variance explained in the first PCA component can be perceived, which suggests that information from several frequency channels is now better represented. This can be clearly seen in table 2, that shows that coefficient construction of the first PCA component – before SUVREL – has positive contribution from few MFCCs. After SUVREL's preprocessing, a positive contribution can be observed from all MFCCs which suggests that this new PCA score space does not "waste" possibly useful information or let it be scattered in several PCA's components. Since more than 90% of all variance explained is compacted amongst the first three components, less computational effort to make the classification tends to be required if SUVREL is adopted as a preprocessing tool.

Table 1: Percentage of the total variance explained by first three PCA components. Data from 100 male subjects.

	original quality	SUVREL	SNR = 25 dB	SUVREL
PCA1	51.71	96.24	60.06	98.17
PCA2	13.79	3.00	12.71	1.70
PCA3	7.68	0.63	6.39	0.08
Total	73.17	99.87	79.15	99.96

Figures 1 and 2 present crossed plots of PCA's scores between the first three components axes. While in fig. 1, produced

Table 2: Contribution of each MFCC to the first PCA component. Data from 100 male subjects.

MFCC	original	SUVREL	MFCC	original	SUVREL
2	0.892	0.512	23	-0.010	0.141
3	0.050	0.184	24	0.000	0.145
4	-0.267	0.042	25	0.002	0.146
5	-0.103	0.107	26	-0.009	0.142
6	-0.127	0.094	27	-0.005	0.144
7	-0.202	0.066	28	-0.002	0.144
8	-0.168	0.078	29	-0.003	0.144
9	-0.071	0.116	30	-0.005	0.144
10	-0.093	0.107	31	-0.003	0.144
11	-0.094	0.107	32	-0.003	0.144
12	-0.048	0.124	33	-0.001	0.145
13	-0.020	0.136	34	-0.005	0.143
14	-0.034	0.131	35	-0.004	0.144
15	-0.033	0.133	36	-0.002	0.145
16	-0.026	0.135	37	0.000	0.146
17	-0.040	0.129	38	0.000	0.145
18	-0.054	0.124	39	0.000	0.146
19	-0.021	0.137	40	0.000	0.146
20	-0.014	0.139	41	-0.001	0.145
21	-0.012	0.141	42	-0.001	0.145
22	-0.014	0.140	43	0.000	0.145

with original data, any clear tendency towards organization is not noticed, in fig. 2, presenting data preprocessed with SUVREL, it is possible to note an obvious aligning tendency with first component axis.

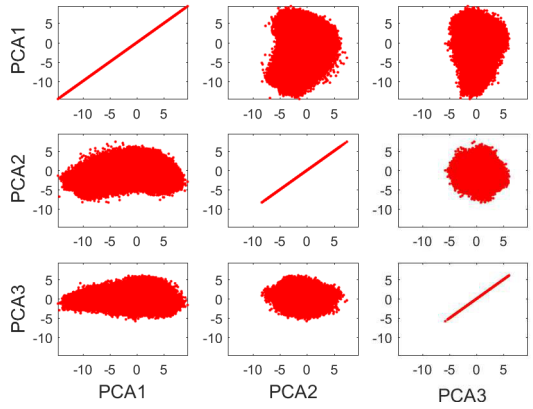


Figure 1: Crossed plots of PCA's scores along first three components axes: any clear tendency towards organization is not noticed.

To have a better visualization of subjects classification, fig. 3 shows each individual as an ellipsoid centered at the PCA's score mean value and stretched proportionally to a fourth of its standard deviation along each PCA component axis, after preprocessing with SUVREL.

Volume intersections (using complete standard deviation) were determined via Monte Carlo and each percentage, with respect to each ellipsoid's own volume, were plotted: fig. 4 is for the original data and fig. 5, after SUVREL. In fig. 5, a percentual decrease of 20.44 ± 0.05 % in the overall volume intersection in comparison with fig. 4 was observed, which can

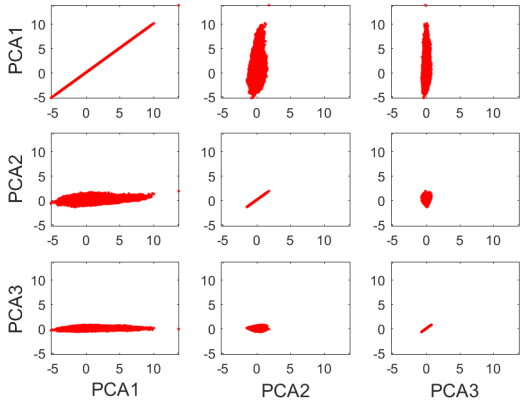


Figure 2: Crossed plots of PCA's scores along first three components axes: shows a clear aligning tendency with first component axis.

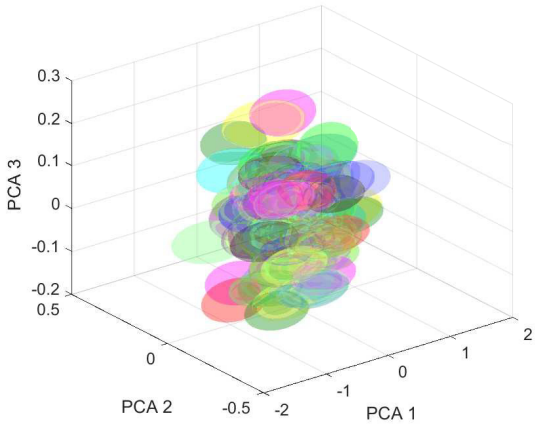


Figure 3: Ellipsoids centered at their PCA's score's mean values and stretched proportionally to a fourth of their standard deviations along each PCA component axis after preprocessing with SUVREL. Regardless of the repeated colors, each ellipsoid represents an individual.

lead to a better subject classification.

Similar analysis were performed with data set in which a white noise (SNR $= 25$ dB) was added. A percentual decrease of 16.22 ± 0.05 % in the overall volume intersection was observed.

Also, MFCC data was obtained from 30 seconds sample utterances (as described in sec. 3) from each one of 100 female subjects. Table 3 shows the increase in the total variance explained after SUVREL preprocessing. In the PCA's scores analysis, the percentual intersections of the ellipsoid's volumes resulted in a decrease of 6.65 ± 0.05 % in the original quality case and of 8.15 ± 0.05 % in the case in which white noise was added.

5. Conclusions

The present work shows the results of a research on the practical application of Supervised Variational Relevance Learning

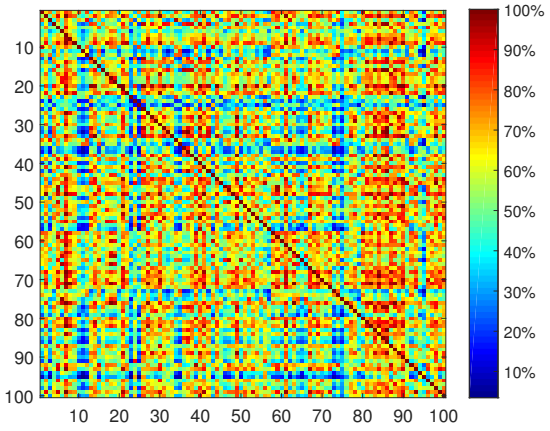


Figure 4: Ellipsoid's volumes intersections represented as temperatures. Each square represents the percentual of volume shared between ellipsoids horizontally numbered with vertically numbered.

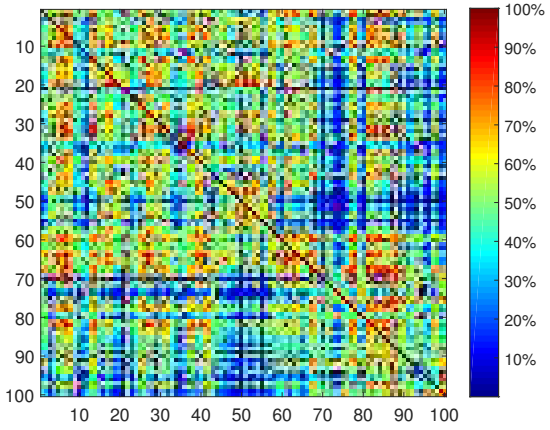


Figure 5: Ellipsoid's volumes intersections represented as temperatures. Each square represents the percentual of volume shared between ellipsoids horizontally numbered with vertically numbered. There is a percentual decrease in the overall volume intersection of 20.44 ± 0.05 % in comparison with fig. 4.

Table 3: Percentage of the total variance explained by first three PCA components. Data from 100 female subjects.

	original quality	SUVREL	SNR = 25 dB	SUVREL
PCA1	48.41	95.68	55.88	98.53
PCA2	14.53	3.54	14.20	1.25
PCA3	8.15	0.63	6.15	0.15
Total	71.09	99.84	76.22	99.93

(SUVREL) to voice comparison. Results show that data preprocessed with SUVREL lead to considerably better resolution of a classifier such as Principal Component Analysis.

Since this article is an introductory survey, further investigation is needed to determine how SUVREL preprocessing affects voice comparison in other situations such as: 1) sample frequency rates less than 44 100 Hz, 2) extraction of other features, instead of MFCC, 3) addition of louder noise, 4) different number of subjects, 5) comparison with results for just the principal diagonal calculation of $g_{\mu\nu}$ (eq. 13) etc.

6. Acknowledgements

The authors thank CAPES – Brazilian Federal Agency for Support and Evaluation of Graduate Education within the Ministry of Education of Brazil – for the travel financial support. Special thanks are due to: Professor Geoffrey S. Morrison, for making available the voice recordings database used in this work and for his constant assistance; Professor Nestor Caticha, IF-USP, and Jonatas Cesar, IB-USP, for the discussions and insights; Andre Borges and Cassia Nozawa for their always available ears.

7. References

[1] A. Bellet, A. Habrard, and M. Sebban, "A Survey on Metric Learning for Feature Vectors and Structured Data," *CoRR*, vol. abs/1306.6, p. 57, 2013. [Online]. Available: http://arxiv.org/abs/1306.6709

[2] M. Boareto, J. Cesar, V. B. P. Leite, and N. Caticha, "Supervised Variational Relevance Learning, an analytic geometric feature selection with applications to omic datasets," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 12, no. 3, pp. 705–711, 2015. [Online]. Available: https://doi.org/10.1109/TCBB.2014.2377750

[3] B. Kulis, "Metric Learning: A Survey," *Foundations and Trends® in Machine Learning*, vol. 5, no. 4, pp. 287–364, 2013. [Online]. Available: http://ieeexplore.ieee.org/xpl/articleDetails.jsp?arnumber=8186753

[4] G. Morrison, C. Zhang, E. Enzinger, F. Ochoa, D. Bleach, M. Johnson, B. Folkes, S. De Souza, N. Cummins, and D. Chow, "Forensic database of voice recordings of 500+ Australian English speakers," 2015. [Online]. Available: http://databases.forensic-voice-comparison.net/

[5] G. S. Morrison, 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*, vol. 44, no. 2, pp. 155–167, 2012. [Online]. Available: http://dx.doi.org/10.1080/00450618.2011.630412

[6] X. Huang, A. Acero, and H. wuen Hon, *Spoken Language Processing - a guide to theory, algorithm, and system development*, 1st ed., J. Bonnell, Ed. Prentice Hall PTR, 2001.

[7] J. F. Hair Jr., W. C. Black, B. J. Babin, and R. E. Anderson, *Multivariate Data Analysis*, 7th ed. Pearson, 2009.

Performance comparison of a number of procedures for computing strength of speech evidence in forensic voice comparison

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

[1] Department of Electrical and Computer Engineering, The University of Auckland, New Zealand
[2] Audio Research Group, Singapore University of Technology & Design, Singapore

h.mehdinezhad@auckland.ac.nz, bj.guillemin@auckland.ac.nz, balamurali_bt@sutd.edu.sg

Abstract

The performance of four procedures for calculating Likelihood Ratios in a Forensic Voice Comparison are compared in this paper. Three of these are existing approaches, namely Gaussian Mixture Model-Universal Background Model, Identity Vector, and Principal Component Analysis Kernel Likelihood Ratio. The fourth, Independent Component Analysis Kernel Likelihood Ratio is a new approach modelled on the latter. As appropriate, experiments were conducted on both tokenized and data-stream based data using Mel-Frequency Cepstral Coefficients as the speech features. In terms of both accuracy and reliability, the new approach is shown to have comparable performance to the existing methods mentioned.

Index Terms: FVC, MFCCs, GMM-UBM, i-vector, PCAKLR, ICAKLR.

1. Introduction

The Bayesian Likelihood Ratio (LR) framework is being increasingly used for evaluating the strength of speech evidence in a Forensic Voice Comparison (FVC) [1]. Among the different procedures available for calculating LRs, Gaussian Mixture Model-Universal Background Model (GMM-UBM) [2] is perhaps the most widely used. This procedure was primarily designed for data-stream-based analysis, but it has also been applied to tokenized data with good results [3]. I-vector analysis is a popular framework for use in speaker verification [4] and some researchers have begun investigating its appropriateness for FVC as well [5]. Similar to GMM-UBM, i-vector is primarily designed for data-stream based analysis. Principal Component Analysis Kernel Likelihood Ratio (PCAKLR) [6, 7] is a relatively new approach for computing LRs that is primarily designed for token-based analysis. In this paper we present a new approach for computing LRs for tokenized data called Independent Component Analysis Kernel Likelihood Ratio (ICAKLR). This is closely modelled on PCAKLR but uses Independent Component Analysis (ICA) rather than Principal Component Analysis (PCA) for producing uncorrelated feature sets. Multivariate Kernel Density (MVKD) and Multivariate Normal (MVN) models [8] have also been used by researchers for calculating LRs but are not included in our comparison experiments.

Performance comparisons of a number of these probabilistic methods have been previously reported. For instance, comparison of MVKD and GMM-UBM when applied to tokenized data was reported in [3] and it was shown that the latter outperformed the former both in terms of accuracy and reliability. The performance of MVKD and PCAKLR for tokenized data was compared in [7] and it was reported that for a large number of input parameters, PCAKLR outperformed MVKD in terms of accuracy. In [9] performance of MVKD, PCAKLR and MVN models were assessed, both before and after fusing with a baseline GMM-UBM system. It was reported that MVKD provided the highest improvement in accuracy to the baseline system, however reliability was compromised significantly. PCAKLR resulted in a marginal improvement in accuracy, but a sizable improvement in reliability. MVN showed only minor improvements in both accuracy and reliability. This paper compares the performance of four probabilistic procedures - three existing, namely GMM-UBM, PCAKLR and i-vector, and a new approach, namely ICAKLR.

Section 2 of this paper provides an overview of the LR framework, followed by a brief discussion of GMM-UBM, i-vector, PCAKLR and ICAKLR. Section 3 describes our experimental procedures for comparing their performance when applied to both data-stream-based and tokenized speech data. Separate experiments have been conducted for speech sampled at both 8 kHz and 32 kHz. We appreciate that a sampling frequency of 32 kHz is not forensically realistic, but we were interested to investigate the extent to which including higher frequency information might improve FVC performance. 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

Mathematically the LR 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 measures 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, values significantly less than one support the defense hypothesis, and 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, i-vector, PCAKLR and ICAKLR

2.2.1. GMM-UBM

GMM-UBM [2, 10] is a commonly used procedure in both automatic speaker recognition and FVC. With this procedure a single background model, typically referred to as a Universal Background Model (UBM), is created initially. It requires a large amount of data for its creation and is trained using all data pooled across all speakers in the relevant background population. The probability density function for the UBM is estimated using Gaussian Mixture Models (GMMs) and the Expectation Maximization (EM) algorithm is used for finding an optimal fit to the data. The Suspect model is then created from this UBM by adapting it towards a better fit for the Suspect speech data, this adaptation being achieved via a Maximum A Posterior (MAP) procedure. The Offender data is compared against both the Suspect and Background models using Equation (1) to produce a score. Once calibrated, this score becomes an LR [11]).

2.2.2. i-vector

In the first stage of the i-vector procedure a speaker-session independent UBM is trained in the same way as for GMM-UBM. Baum-Welch statistics are then extracted from this [12, 13]. Factor Analysis (FA) is used to represent a new low-dimensional subspace called Total Variability (TV) using EM. Identity vectors (i.e., i-vectors) are then estimated from this [12-14]. Unlike GMM-UBM in which acoustic feature vectors represent the test segments, the i-vector framework test segments are represented using i-vectors, with Linear Discriminant Analysis (LDA) being used to reduce their dimensionality [15]. A Generative Factor Analysis approach called the Probabilistic LDA (PLDA) is used to model the i-vectors [16]. Finally, scores and then LRs are computed from these.

2.2.3. PCAKLR

In PCAKLR the speech features are transformed into a new set of uncorrelated features using Principal Component Analysis (PCA) [6]. Individual scores are computed for each of these using Univariate Kernel Density (UKD) analysis [17]. Since these transformed parameters are uncorrelated, a final score can be determined by multiplying the individual scores. Calibrating this result produces an LR.

2.2.4. ICAKLR

ICAKLR is very similar to PCAKLR except that ICA rather than PCA is used to transform the speech data into a new set of independent (i.e., uncorrelated) features [18]. The motivation for using ICA is that it has been used in speech recognition applications [19]. We were therefore interested to see how it might perform in the FVC arena. Scores and LRs can then be calculated from this set of independent features in the same way as for PCAKLR.

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) [20]. Accuracy indicates the closeness of the obtained result to the true value of the output. The Log-Likelihood Ratio Cost (C_{llr}) [21] 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. The Credible Interval (CI) [21] 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 which represent the cumulative proportion of LLR values for both same-speaker and different-speaker comparisons [22]. In these plots (see Fig. 1) the solid blue and solid red curves are the same-speaker and different-speaker comparison results, respectively. Since positive LLR values support the same-origin hypothesis and negative values support the different-origin 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 result 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 (Extended Multi Modal Verification for Teleservices and Security) speech database [23] was used in this investigation. This multi-modal database contains read speech digitized at 16 bits, sampled at 32 kHz, and the background noise level is low. This data was then down-sampled for use in our 8 kHz experiments. The language in the database is English with predominantly a Southern British accent. It contains four recording sessions of 295 subjects (156 male, 139 female) 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”.

Given that the XM2VTS database contains recordings of read speech and the background noise level is low, it is acknowledged that it is not very forensically realistic [24]. However, in support of its use in our experiments, it does include a large number of speakers with similar accent as well as multiple non-contemporaneous recordings, both aspects being highly important in the FVC arena. Of the 156 male speakers, only 130 were used for this study. 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 [24] for the rationale behind discarding recordings on the basis of dissimilar accent).

For our data-stream based experiments, the whole utterance after removing any silence segments was used. For the tokenized data experiments, two diphthongs /aɪ/ and /eɪ/ and one monophthong /i:/ extracted from the words “nine”, “eight” and “three”, respectively, were used.

Mel-Frequency Cepstral Coefficients (MFCCs) were used for this investigation [19-21]. Our tokenized data experiments used 14 MFCCs, while our data-stream-based experiments used 14 MFCCs, 14 Deltas and 14 Delta-Deltas.

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 [11]). 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 [25].

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 [11]. 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. CI for both same-speaker and different-speaker comparison results were computed using the procedure outlined in [16].

4. Results

Tippett plots of our experiments are in Figure 1 and mean C_{llr} and 95% CI are in Table 1. We consider first the results for 8 kHz speech data, then compare results between 8 kHz and 32 kHz data.

For the 8 kHz token-based experiments, and considering first FVC accuracy, PCAKLR, ICAKLR and GMM-UBM (Fig. 1 a-c) have given very similar performance, with GMM-UBM marginally outperforming the other two, this being linked to its comparatively lower number of different-speaker misclassifications. The performance of i-vector (Fig. 1 d) in terms of accuracy is slightly worse than the other three, this being caused by a slightly larger number of different-speaker misclassifications. As expected, the larger amount of information available with data-stream-based analysis gives rise to significantly improved FVC accuracy. Comparing Fig. 1 (c) with Fig. 1 (i) for GMM-UBM shows that this increased data has resulted in a very low number of misclassifications, both same-speaker and different-speaker. The same observation is true for i-vector (compare Fig. 1 (d) with Fig. 1 (j)).

Again for the 8 kHz token-based experiments, and considering now FVC reliability, it is interesting to note that GMM-UBM is worse than the other three procedures in this regard. Comparing Fig. 1 (a-d) it is clear that this is due to the reliability of GMM-UBM in respect to same-speaker comparisons being worse than for the other three procedures. Compared to token-based analysis, data-stream-based analysis also gives significantly improved performance in terms of reliability (compare Fig. 1(c) with Fig. 1 (i) and Fig. 1 (d) with Fig. 1 (j)).

Now to the question of the impact of sampling frequency on FVC performance, though there is definitely an improvement with higher sampling frequency, interestingly this is not that significant (compare for example Fig. 1 a-d with Fig. 1 e-h for token-based analysis and Fig. 1 (i) with Fig. 1 (k) for data-stream-based analysis).

5. Conclusions

An FVC comparison performance of three existing approaches for calculating LRs, namely GMM-UBM, i-vector and PCAKLR, and one new approach called ICAKLR, has been presented in this paper. The first two methods were applied to both tokenized and data-stream-based data, whereas the latter two were applied solely to tokenized data. Separate experiments were conducted using both 8 kHz and 32 kHz speech data in order to investigate the impact of sampling frequency on FVC performance.

Firstly, comparing PCAKLR with the new procedure presented, namely ICAKLR, ICAKLR has only marginally outperformed PCAKLR. One could conclude from this that the de-correlating performance of ICA is very similar to that of PCA in this application. For token-based analysis, i-vector does not perform as well as the other three. As expected, if sufficient data is available in order to conduct a data-stream-based analysis, significantly improved FVC performance results. As far as sampling frequency is concerned, though a higher sampling frequency does result in improved FVC performance, this is not that significant.

6. References

[1] Morrison, G.S., *Likelihood-ratio forensic voice comparison using parametric representations of the formant trajectories of diphthongs.* the Journal of the Acoustical Society of America, 2009. **125**: p. 2387.

[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] 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.

[4] Kanagasundaram, A., *Speaker verification using I-vector features*. 2014, Queensland University of Technology.

[5] Huang, C.C., J. Epps, and T. Thiruvaran, *An investigation of supervector regression for forensic voice comparison on small data.* EURASIP Journal on Audio, Speech, and Music Processing, 2015. **2015**(1): p. 7.

[6] Nair, B., E. Alzqhoul, and B.J. Guillemin, *Determination of likelihood ratios for forensic voice comparison using Principal Component Analysis.* International Journal of Speech, Language & the Law, 2014. **21**(1).

[7] 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.

[8] Aitken, C.G. and D. Lucy, *Evaluation of trace evidence in the form of multivariate data.* Journal of the Royal Statistical Society: Series C (Applied Statistics), 2004. **53**(1): p. 109-122.

[9] Enzinger, E. *Likelihood Ratio Calculation in Acoustic-Phonetic Forensic Voice Comparison: Comparison of Three Statistical Modelling Approaches*. in *INTERSPEECH*. 2016.

[10] Reynolds, D., *Gaussian Mixture Models,* . Encyclopedia of Biometric Recognition, Springer, 2008.

[11] 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.

[12] Dehak, N., et al., *Front-end factor analysis for speaker verification.* IEEE Transactions on Audio, Speech, and Language Processing, 2011. **19**(4): p. 788-798.

[13] Kenny, P. *A small footprint i-vector extractor*. in *Odyssey 2012-The Speaker and Language Recognition Workshop*. 2012.

[14] Matrouf, D., et al. *A straightforward and efficient implementation of the factor analysis model for speaker verification*. in *Eighth Annual Conference of the International Speech Communication Association*. 2007.

[15] Prince, S.J. and J.H. Elder. *Probabilistic linear discriminant analysis for inferences about identity*. in *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*. 2007. IEEE.

[16] Kenny, P. *Bayesian speaker verification with heavy-tailed priors*. in *Odyssey*. 2010.

[17] Lindley, D., *A problem in forensic science.* Biometrika, 1977. **64**(2): p. 207-213.

[18] Stone, J.V., *Independent component analysis: a tutorial introduction*. 2004: MIT press.

[19] Chien, J.-T. and B.-C. Chen, *A new independent component analysis for speech recognition and separation.* IEEE transactions on audio, speech, and language processing, 2006. **14**(4): p. 1245-1254.

[20] Morrison, G.S., *Forensic voice comparison and the paradigm shift.* Science & Justice, 2009. **49**(4): p. 298-308.

[21] Morrison, G.S., *Measuring the validity and reliability of forensic likelihood-ratio systems.* Science & Justice, 2011. **51**(3): p. 91-98.

[22] 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.

[23] Messer, K., et al. *XM2VTSDB: The extended M2VTS database*. in *Second international conference on audio and video-based biometric person authentication*. 1999.

[24] 3GPP2-S0018-D, *S0018-D, Minimum Performance Specification for the Enhanced Variable Rate Codec, Speech Service Options 3, 68, 70, and 73 for Wideband Spread Spectrum Digital Systems. Retrieved on 2 June 2013, last retrieved from http://www.3gpp2.org/*. 2012.

[25] 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*, in *SST 2016, Parramatta, Australia*.

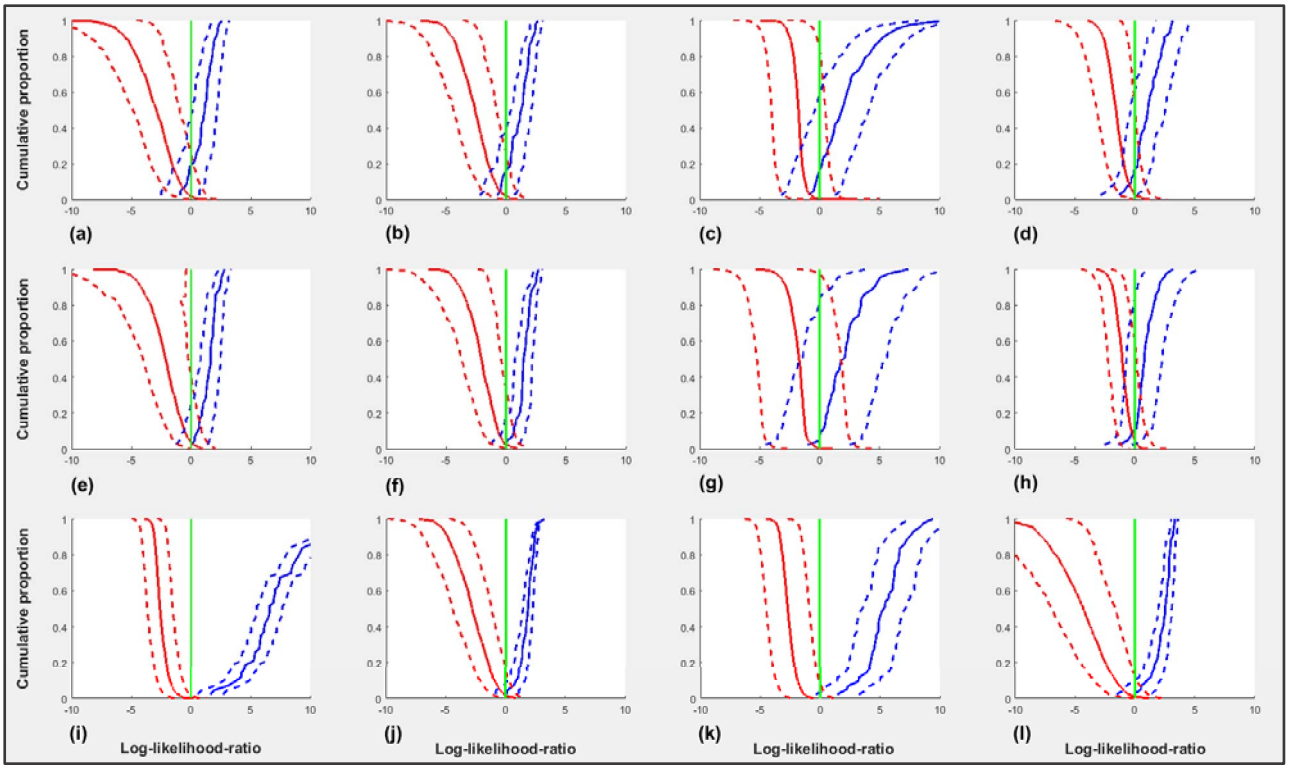


Figure 1. *Tippett plots showing the results of our investigation. The solid lines indicate LLR values and dashed lines on either side of these represent estimates of the 95% CI values. PCAKLR for tokenized data: (a,e); ICAKLR for tokenized data: (b,f); GMM-UBM for tokenized data: (c,g); i-vector for tokenized data: (d,h); GMM-UBM for data-stream data: (i,k); i-vector for data-stream data: (j,l). Figs. a-d and i-j are for 8 kHz speech and Figs. e-h and k-l are for 32 kHz speech.*

Table 1. *Mean C_{llr} and 95% CI for experiments*

		Tokenized Data				Streamed Data	
		PCAKLR	ICAKLR	GMM-UBM	i-vector	GMM_UBM	i-vector
8 kHz	C_{llr}	0.269	0.233	**0.214**	0.328	**0.008**	0.075
	CI	2.128	1.852	2.245	**1.645**	**1.069**	1.555
32 kHz	C_{llr}	0.129	**0.115**	0.139	0.352	**0.006**	0.050
	CI	2.193	1.759	3.515	**1.199**	**1.796**	2.698

Emotion Recognition Using Intrasegmental Features of Continuous Speech

Li Tian, Catherine Inez Watson
Department of Electrical and Computer Engineering, University of Auckland, New Zealand
tli725@aucklanduni.ac.nz, c.watson@auckland.ac.nz

Abstract

This paper proposed a new emotion recognition system using intrasegmental features, extracted from long monophthongs in continuous speech. 36 vocal tract features and 11 glottal source features were initially extracted and an optimal subset was selected using Maximum Relevance Minimal Redundancy Backward Wrapping (MRMRBW). A newly constructed JL corpus was used to evaluate the system performance. Five different classifiers were considered. By using the optimal classifier, we achieved recognition accuracy of 70.5% regardless of vowel types for five different emotions.

Index Terms: emotion recognition, intrasegmental features, glottal source, feature selection, IAIF

1. Introduction

Advanced human to machine communication systems seek not only explicit semantic exchange but also the transmission of the vocal expression of emotion, which carries a person's hidden intent, motive and physiology state. A thorough understanding of a spoken utterance can only be achieved by recognizing the incorporated emotion. Speech emotion is detected via a combination of features at all three principal levels of speech abstraction: suprasegmental, segmental and intrasegmental [1]. Most of the previous studies [2] focused on the suprasegmental and segmental level. Various acoustic features such as fundamental frequency (F0), intensity and speech rate are identified as effective emotion descriptors. However, the significance of intrasegmental features in emotion recognition has not been widely investigated. The intrasegmental characteristics reflect the vocal behavior down to the specific phoneme level [5]. Since emotions can be differentiated in segments of running speech as short as 60ms by humans [3], more insight should be provided into the features for these phonetic segments. During the intrasegmental period, voice change affected by vocal expression of emotions can be implemented through the manipulation of vocal tract shape and vocal cords vibration pattern. While the vocal tract can be adequately represented by its spectral features such as MFCCs [2], parametrizing the glottal source is often problematic [4]. One of the issues is the waveform's non-periodicity. This can be partially solved by considering monophthong tokens. A few studies [5] found the evidence of the emotion discriminative power of vowels in a sentence using their prosody or glottal source features. However, knowledge of the topic is still far from conclusive. Hence, this paper examined both vocal tract and glottal source features extracted from 4 frequently encountered long monophthongs of New Zealand English and developed an automatic emotion recognition system to evaluate their performance. The corpus used in the study is discussed in Section 2, followed by the extraction methodology of proposed 47 intrasegmental features and the resulting feature selection in Section 3 and Section 4 respectively. Section 5 describes the emotion classification models used and their performance results are discussed in Section 6.

2. Emotional Corpus

To meet the research goal, a speech corpus (called JL [6]) with strictly-guided simulated emotions was developed. It has many of the long vowels (/a:/, /i:/, /o:/ and /u:/), equally distributed in the words of 15 semantically neutral sentences. Short sentences with 4-7 syllables were chosen to lower the risk of deviation from the target emotion when the speakers went throughout each of the sentences. Four professional New Zealand English speakers (two male and two female) produced the speech with ten target emotions (5 primary and 5 secondary ones). The primary emotions are happy, angry, neutral, sad, excited and the secondary emotions are enthusiastic, apologetic, pensive, worried, and anxious. The emotions were elicited and calibrated by presenting the same strong emotion related backgrounds to the speakers during the recording. The recording was done in a soundproofed room. An AKG C460B microphone and a Roland Octa-Capture pre-amplifier (set to 25 dB) were used to collect the speech data, sampled at 44.1kHz and stored as 16-bit numbers. In summary, there are 4 (speakers) × [5 (primary emotions) +5 (secondary emotions)] × 2 (repetitions) × 15 (sentences) × 2 (sessions), making a total of 2400 sentences.

To enable performance comparison with other existing corpora [2], the primary emotion subset of the JL corpus was used for this study. The audio dataset was automatically labeled at the word and phonetic levels using the Munich Automatic

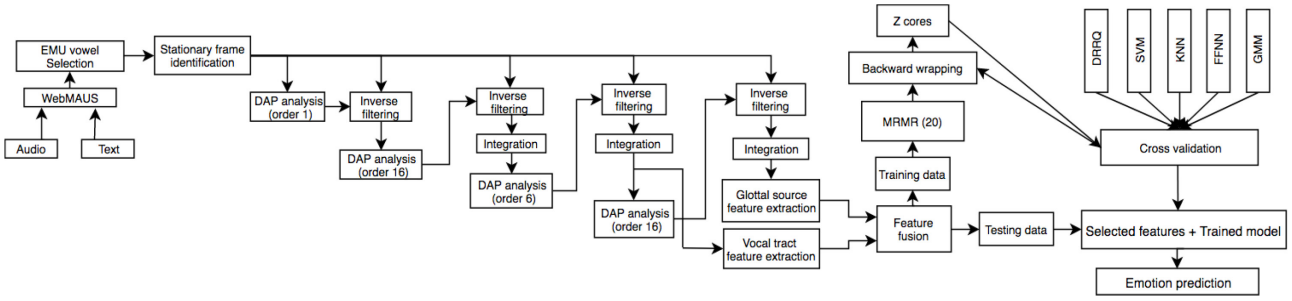


Figure 1: *Overview of the emotion recognition system design*

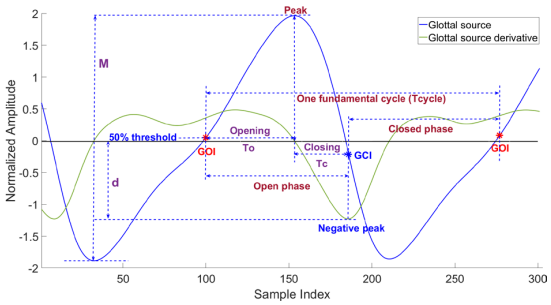


Figure 2: *A segment of the glottal source signal*

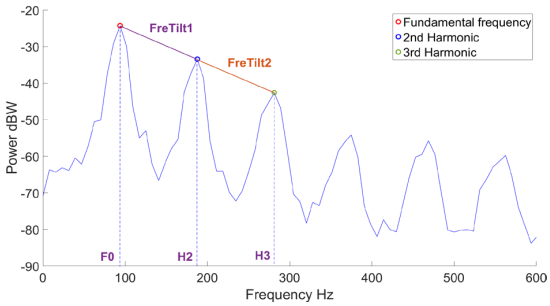


Figure 3: *Glottal source spectral tilt*

Web Segmentation System, webMAUS [7], followed by the conversion to an EMU formatted database [8]. Additional manual corrections were done to adjust the wrongly marked vowel boundaries. Then, each vowel segment was extracted from the vowel onset to vowel offset. The first and last 5% portion of the extracted vowels were neglected for further feature extraction to exclude the possible coarticulation from their adjacent phonemes. The final dataset has 2920 tokens (730 /a:/, 730 /i:/, 730 /o:/ ,730 /u:/), each with an emotion class of their corresponding sentences. The tokens were permutated to guarantee a vowel type independent emotion recognition. These vowels were assumed to be appropriate indicators of the emotion tag for the whole expression, as per [5].

3. Extraction of intrasegmental features

Iterative Adaptive Inverse Filtering-IAIF [9] was applied to decompose all the speech waveforms into vocal tract part and glottal source part. Both parts were used to derive emotion discriminative information here. IAIF is a two-stage iteration process based on the principle of discrete all pole modeling (DAP) [10]. It recursively estimates the vocal tract model for every Hanning windowed 20ms analysis frame. After obtaining a re-estimated vocal tract signal, a final glottal source can be calculated by inverse filtering this vocal tract and lip radiation model from the original speech (i.e., the vowel segments). Details can be seen in Figure 1. Whilst an optimal glottal waveform can be produced by tuning the order of DAP estimation and lip radiation coefficients for each recording, for automatic processing, they were set to be constant, 16 and 0.99 respectively, based on pilot experiments [11].

3.1. Parameterization of the vocal tract and glottal source

For the glottal source signal, a sliding rectangular window, shifting from beginning to end of each vowel segment by 1ms was used to identify stationary analysis frame. 7 time domain and 4 frequency domain features were extracted from this frame. The preset window length and identification threshold were adjusted according to the vowel length [15]. Time domain features used are Glottal Source Zero Cross Rate (GSZCR), mean Open Quotient (mOQ), Open Quotient Perturbation (OQP), mean Speed Quotient (mSQ), Speed Quotient Perturbation (SQP), mean Normalized Amplitude Quotient (mNAP) and Normalized Amplitude Quotient Perturbation (NAQP). Prior to these feature extractions, some critical time instances were first located in each segment, seen in Figure 2. In accordance with the LF model [12] and sub50 open quotient definition [13] the Glottal Opening Instance (GOI) and Glottal Closing Instance (GCI) in each cycle were determined as the instances when the glottal source signal increasingly crossed a 50% peak-valley amplitude threshold and its corresponding derivative reached the minimal flow respectively. The peak point between each GOI and GCI pair represents the maximal glottal opening when the glottal air flow at a maximum. The Open Quotient (OQ) is the time ratio of when the glottal folds are open ($T_o + T_c$) to the corresponding duration of the fundamental cycle (T_{cycle}). The Speed Quotient (SQ) is the time ratio of the opening phase (T_o) over the closing phase (T_c). The Normalized Amplitude Quotient (NAQ) [14] is the ratio of the glottal flow cycle peak (M) to the product of the amplitude of its first derivative's minimum (d) and the fundamental cycle. The mean and perturbation of OQ, SQ and NPQ within the analysis frame were extracted as the features. Unlike the standard deviation, which quantifies the overall variance of the dataset, the perturbation was reported to be a more robust reflection of the behaviors of the vocal folds [15]. Based on jitter and shimmer calculations, perturbation measures the average parameter difference between cycles divided by the mean for each cycle. It is given by:

$$P_x = \frac{100 * \frac{1}{N-1}\sum_{n=1}^{N-1}|x_{n+1} - x_n|}{\frac{1}{N}\sum_{n=1}^{N} x_n} \quad (1)$$

where x can be OQ, SQ or NPQ, and N is the number of cycles in the analysis frame. Within the same analysis frame, GSZCR is also derived from calculating zero crossing rate (ZCR).

Welch's power spectral density estimate [16] was used to convert the glottal source signal to the frequency domain. This method reduces the random noise, caused by both irregular glottal closures and relatively short sample lengths. Various glottal source frequency tilt values in the region between 0 to 3700 Hz have been found effective in depression disorder detection [17]. Inspired by their results, this study also investigated spectral tilt for emotion classification. Figure 3 depicts an example of a glottal source spectrum calculated from an /a:/ token. The spectral peaks can be easily spotted, and two lines are well fitted to the first three peaks. The three frequency features depending on these peaks are defined as:

$$FreTilt1 = \frac{P(F0) - P(H2)}{F0 - H2} \quad (2)$$

$$FreTilt2 = \frac{P(H2) - P(H3)}{H2 - H3} \quad (3)$$

$$HighFre = \frac{P(f > H3)}{P(f > 0)} \quad (4)$$

where P denotes spectral power, and $F0$, $H2$ and $H3$ represent the fundamental frequency, the second and third harmonics respectively. The features *FreTilt1* and *FreTilt2* describe the two-step power dropping rate from the fundamental frequency to the third harmonic while $HighFre$ specifies the cumulative power impact of those high frequency ($>H3$) components.

Additionally, a power based harmonic richness factor (HRF) [18] was extracted, given as:

$$HRF = \frac{\sum_{n\geq 2} P(H_n)}{P(F0)} \quad (5)$$

where $P(H_n)$ is the spectral power of the F0's nth harmonic and $P(F0)$ is the spectral power of the F0.

As for the vocal tract signal, it was assumed to be sufficiently parameterized by 36 spectral features: the first 12 MFCC coefficients, the first 12 delta MFCC coefficients (MFCC first-order derivatives) and the first 12 double delta MFCC coefficients (MFCC second-order derivatives).

4. Feature selection

Not every proposed feature will necessarily be useful for emotion recognition, and many are highly correlated. Therefore, we used a feature selection method to retrieve a subset from the initial set of 47 features. The method we used to optimize the feature set is the combination of maximum relevance minimum redundancy (MRMR) [19] and backward wrapping approaches. MRMR is implemented by selecting a feature subset that has a maximum normalized sum of the relevance of each feature with the category vector, as well as a minimum normalized sum of mutual information between any of its two features. In this study, mutual information difference (MID) scheme [19] was chosen to achieve these two goals simultaneously. If S denotes a set of selected features within the overall feature set Ω and c is the class or category vector (for this study, each element of c is a single vowel's emotion class), then the optimal set of features S_O can be calculated as:

$$S_o = \underset{S\subset\Omega}{argmax} \{\frac{1}{|S|}\sum_{x_i\in S} I(x_i;c) - \frac{1}{|S|^2}\sum_{x_i,x_j\in S} I(x_i,x_j)\} \quad (6)$$

where $|S|$ is the number of features in S, x is the feature vector and I is the mutual information of two arbitrary input vectors, given as:

$$I(x;y) = \iint p(x,y) log \frac{p(x,y)}{p(x)p(y)} dxdy \quad (7)$$

where p denotes the probabilistic density function. In practice, an incremental search was applied to find the near-optimal S_o. For example, suppose we have already found out the optimal S_{n-1} subset with *n-1* features, the next step is to find the *nth* feature from the rest feature set by calculating:

$$x_n = \underset{x_n\in\Omega-S_{n-1}}{\text{argmax}} \{I(x_n;c) - \frac{1}{n-1}\sum_{x_m\in S_{n-1}} I(x_m;x_n)\} \quad (8)$$

And the first feature is selected as the one that has the maximal mutual information with the class vector c. This incremental search stops when we reach the desired feature set size. This preset size was heuristically decided as 20, considering 47 original extracted features.

Figure 4: *Emotion recognition confusion matrix*

The backward wrapping followed the MRMR process (see in Figure 1). It started with the MRMR selected feature set and removed the poor features in a backward manner. For each round, one of the existing features was excluded such that the remaining ones produced the highest recognition performance. Such performance was measured as the mean of 5 folds cross-validation accuracy scores using the given classifier. To avoid being trapped in local minima, the round continued until two features were left. The highest performance of each round formed a sequential vector. By sorting this vector, the best-performing round can be identified, where the optimal feature subset can be found accordingly. In summary, the whole MRMRBW process selected the least number of features that are both emotion-general and classifier-adaptive. All of these features were centered and normalized using their z-scores before entering the emotion classifiers.

5. Emotion classification

Five types of classifiers including Double Round Robin Quadratic Model (DRRQM) [11], Feed-Forward Neural Network (FFNN) [20], Support Vector Machines (SVM) [21], K-Nearest Neighbor (KNN) and Gaussian Mixed Model (GMM) have been used for this study. To avoid overfitting, for each type of the classifiers, a triple nested cross-validation scheme was used. The outer 10-fold cross-validation was used to evaluate the model performance. For each round of the outer cross-validation, the optimal feature subset was selected using the MRMRBW approach on the partitioned training data. This feature selection process further included an intermediate 5-fold cross-validation as the backward wrapping treated the average accuracy of this cross-validation as its feature selection metric. Using the optimal features, a model was re-trained by combining all the partitioned training data. The partitioned testing data then chose the same selected features as this model and was predicted afterwards. For each round of the intermediate cross-validation, another 5-fold inner cross-validation was used on the partitioned sub-training data to tune the hyperparameters of the given classifier prior to the model refitting. If no hyperparameters exist for the given classifier, this inner cross-validation is simply skipped. While no hyperparameters are required for DRRQM, which is the fusion of double round Robin technique [11] and quadratic classifier, the number of closest neighbors and Gaussian components were regarded as hyperparameters for KNN and GMM respectively. LIBSVM [22] package with one versus all (OvA) scheme was used to train the SVM classifier. The choice of the kernel function (radial basis function (RBF) kernel or polynomial kernel), the order of polynomial if polynomial kernel selected, the box constrain were all regarded as hyperparameters for this classifier. Their optimal values or choices were exhaustively

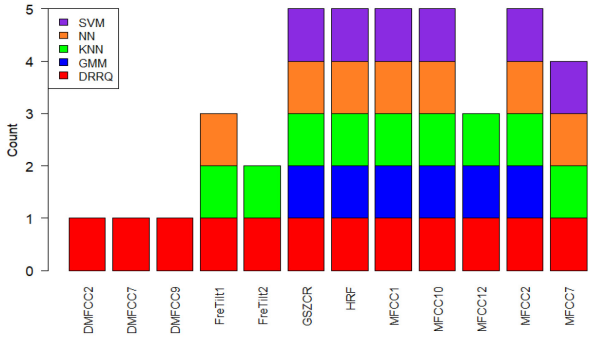


Figure 5: *Generally best features selected for the 5 different classifiers*

searched from the possible ranges [22]. Concerning the FFNN, the number of hidden layers and neurons on each were regarded as its hyperparameters. Based on the actual size of the samples and some empirically derived rules-of-thumb [23], 2 and 40 were suggested as their tuning range maximum respectively.

6. Result and Discussion

For different classifiers, the outer cross-validation predicted emotions against the actual emotions were investigated in the form of confusion matrix. It was found that SVM outperforms others in terms of the overall accuracy. Due to the page limit, only SVM's confusion matrix is shown in Figure 4. The bold black numbers in each cell indicate the number of testing samples, whose actual emotion on the left side was predicted as the emotion at the bottom side. The last row and column of the confusion matrix (grey cell) show the precision and hit rate for each emotion respectively. The overall hit rate for all emotions is indicated in blue cell. Moreover, for each classifier the commonly selected features using MRMRBW from each round of the outer cross-validation were regarded as its generally best input features. Those features selected for different classifiers are illustrated in Figure 5. It is observed that 2 glottal source features (GSZCR and HRF) and 3 vocal tract features (MFCC1, MFCC2 and MFCC10) are commonly preferred by all classifiers for generating the best classification performance. This finding reveals that on the intrasegmental level both the glottal source and vocal tract can act as suitable indicators of the speech emotion. Looking at the confusion matrices, the high arousal emotions (e.g., angry, excited) are well separated from the low arousal ones (e.g., sad). Some of the most confused pairs are happy-angry, sad-neutral and angry-excited, which are apparently different on the valence level. These misclassifications were extensively reported by other emotional speech studies [5, 24]. It can be noted that comparing with some established models [25] using various segmental and suprasegmental spectral and prosody features on similar corpus Emo-DB [26], our overall hit rate is not quite competitive though. One of the cause is that our model is mixed vowel type based. Reflecting on the features, the emotion difference may be overlaid by the vowel type difference, as different vowel types are pronounced using different vocal shapes. This overlay may also explain why only a few of low-order MFCCs were left as generally best features by different classifiers since the others presumably more robust to identify phonetic content rather than non-verbal voice attributes like emotions. Also, when it is down to the intrasegmental level, more emotion type uncertainty should be expected [3]. Vowel type wised modelling is more likely to achieve a higher classification rate.

7. Conclusion

In this paper, we proposed an intrasegmental feature based emotion recognition system. Reasonable emotion recognition rates were obtained regardless of vowel types for the JL primary emotion subset. The significance of vocal source features in emotion recognition was observed through the feature selection result. This study indicates that the vowel segments alone can appropriately work as the emotion descriptor when the optimal glottal source and vocal tract features are extracted.

8. References

[1] I. Murray, J. Arnott, "Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion", *The JASA journal*, 93(2), pp. 1097-1108, 1993.

[2] C. Anagnostopoulos, et al., "Features and classifiers for emotion recognition from speech: a survey from 2000 to 2011", *Artificial Intelligence Review*, 43(2), 2015

[3] I. Pollack, et al., "Communication of Verbal Modes of Expression", *Language and Speech*, 3(3), pp. 121-130, 1960.

[4] C. Dromey, et al., "Glottal airflow and electroglottographic measures of vocal function at multiple intensities", *Journal of Voice*, 6(1), pp. 44-54, 1992.

[5] M. Airas, P. Alku, "Emotions in Vowel Segments of Continuous Speech", *Phonetica*, 63(1), pp. 26-46, 2006.

[6] J. James, L. Tian, C. Watson, "An Open Source Emotional Speech Corpus for Human Robot Interaction Applications", *Interspeech 2018*.

[7] Kisler, T. and Schiel, F., Sloetjes, H, "Signal processing via web services: the use case WebMAUS", *Digital Humanities, Hamburg, Germany, Hamburg*, pp. 30-34, 2012.

[8] S. Cassidy, J. Harrington, "Multi-level annotation in the Emu speech database management system", *Speech Communication*, vol. 33, no. 1-2, pp. 61-77, 2001.

[9] P. Alku, "Glottal wave analysis with Pitch Synchronous Iterative Adaptive Inverse Filtering", *Speech Communication*, vol. 11, no. 2-3, pp. 109-118, 1992.

[10] A. El-Jaroudi and J. Makhoul, "Discrete all-pole modeling", *IEEE Transactions on Signal Processing*, no. 2, pp. 411-423, 1991.

[11] L.Tian, C. Watson, "Continuous Spoken Emotion Recognition Based on Time-Frequency Features of the Glottal Pulse Signal within Stressed Vowels." In *Australasian International Conference on Speech Science and Technology*, 2016.

[12] Fant, Gunnar. "The LF-model revisited. Transformations and frequency domain analysis." *Speech Trans. Lab. Q. Rep., Royal Inst. of Tech. Stockholm,* 2(3), 1995.

[13] C. Sapienza, E. Stathopoulos and C. Dromey, "Approximations of open quotient and speed quotient from glottal airflow and egg waveforms: Effects of measurement criteria and sound pressure level", *Journal of Voice*, 12(1), pp. 31-43, 1998.

[14] P. Alku, T. Bäckström and E. Vilkman, "Normalized amplitude quotient for parametrization of the glottal flow", *The Journal of the Acoustical Society of America*, 112(2) pp. 701-710, 2002.

[15] S. Bier, C. Watson and C. McCann, "Using the Perturbation of the Contact Quotient of the EGG Waveform to Analyze Age Differences in Adult Speech", *Journal of Voice*, 28(3), 2014.

[16] D. France, R. Shiavi, S. et al., "Acoustical properties of speech as indicators of depression and suicidal risk", *IEEE Transactions on Biomedical Engineering*, vol. 47, no. 7, pp. 829-837, 2000.

[17] E. Moore, et al., "Critical Analysis of the Impact of Glottal Features in the Classification of Clinical Depression in Speech", *IEEE Trans on Biomedical Engineering*, 55(1), 2008.

[18] D. Childers, C. Lee, "Vocal quality factors: Analysis, synthesis, and perception", *The Journal of the Acoustical Society of America*, 90(5), pp. 2394-2410, 1991.

[19] H. Peng, F. Long, et al., "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(8), pp. 1226-1238, 2005.

[20] G. Hinton, L. Deng, D. Yu, G. Dahl, "Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups", IEEE SPM, 29(6), pp. 82-97, 2012.

[21] Corinna Cortes Vladimir Vapnik, "Support-Vector Networks", *Machine learning*, vol. 20, no. 3, pp. 273-297, 1995.

[22] C. Chang and C. Lin, "LIBSVM", *ACM Transactions on Intelligent Systems and Technology,* 2(3), pp. 1-27, 2011.

[23] J. Heaton, *Introduction to neural networks with Java*. St. Louis: Heaton Research, 2005.

[24] M. El Ayadi, M. Kamel and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases", *Pattern Recognition*, 44(3), pp. 572-587, 2011.

[25] S. Krothapalli and S. Koolagudi, "Characterization and recognition of emotions from speech using excitation source information", *Int. J. of Speech Technology*, 16(2), 2012.

[26] Burkhardt, F. et al., "A database of German emotional speech," *Interspeech,* pp. 1517–1520, 2005.

Infant-Directed Speech may not be across-the-board breathy, but has a variable voice quality

Titia Benders[12], Elise Tobin[1], Anita Szakay[12]

[1]Department of Linguistics, Macquarie University;
[2]ARC Centre of Excellence in Cognition and its Disorders
{titia.benders, elise.tobin, anita.szakay}@mq.edu.au

Abstract

This study aimed to characterise the difference between Dutch infant-directed speech (IDS) and adult-directed speech (ADS) in terms of voice quality. After controlling for the effect of *F0* and F1 on the acoustic H1-H2 measurement of voice quality, it was found that Dutch IDS may not be across-the-board breathier than ADS, but is spoken with more variability in its voice quality. These results raise questions about the role of voice quality and variability in parent-child interactions.
Index Terms: Infant-Directed Speech; Voice Quality

1. Introduction

Infant-directed speech (IDS) is the emotional register that speakers cross-culturally adopt when addressing infants [1, 2] and that infants prefer listening to over adult-directed speech (ADS) [3]. It is thought that infants' preference for IDS is driven by the "happy" sound of this register [4].

The acoustic profile of IDS commonly consists of a higher fundamental frequency (*F0*) and larger *F0* range within utterances [5]. An IDS characteristic that is emerging across studies are higher formant frequencies [6, 7]. These are all characteristics of speech with positive valence and high arousal, and thus of "happy" speech [8, 9]. Identifying further acoustic characteristics of IDS can help disentangle whether IDS is "positive", "energetic", or inherently both.

IDS does not only sound across the board "happier" than ADS, it is also acoustically more variable: *F0* changes more from one utterance to the next [10, 11] and the spread of the formant frequencies within vowels is larger [12, 13]. It has recently been observed that even the changes within *F0* contours are more surprising in IDS [14]. This variability suggests that speakers may not adopt a static positive valence in IDS, but adapt in a way that signals high arousal [15]. To begin to understand whether and how acoustic variability contributes to the "happy" sound of IDS, it is critical to routinely document it across acoustic characteristics.

A recently identified acoustic feature of IDS is its different voice quality compared to ADS [16]. The IDS voice quality may be partially due to the raised formant frequencies discussed above, but also to excessive "breathiness" [13]. Breathy voice is one of the three main voice qualities - together with modal and creaky voice. Breathy voice is generally defined as having constant airflow through the glottis during vocal fold vibration, which is commonly caused by an incomplete closure of the vocal folds during the closure phase of the glottal cycle [17, 18]. Rather than representing discrete categories, creakiness, modal voice, and breathiness vary along a continuum of glottal closure [19], with increasing values of open quotient towards the breathy end of the scale. Voice quality and breathiness are interesting with respect to IDS, because breathiness in speech is more strongly associated with valence than arousal [20]. However, there are still several open questions regarding voice quality, and specifically breathiness, in IDS, which the present study aims to address.

Firstly, the one study identifying breathiness in IDS was conducted in Japanese [13]. A breathy voice quality marks politeness in Japanese [21, 22], presumably because it signals intimacy and friendliness [21]. As intimacy and friendliness are likely properties of IDS, this begs the question whether IDS is also breathy in languages where it does not serve such a clear paralinguistic function. A second open question is whether voice quality is more variable in IDS compared to ADS, an issue that was not addressed in [13].

Dutch is an excellent language for addressing these two questions. Firstly, breathiness does not serve a (documented) paralinguistic function in Dutch. Second, Dutch IDS has a higher *F0*, larger *F0* range, and higher formant frequencies than ADS [7, 23]. Yet, the extent of the *F0* related exaggeration in IDS is relatively small compared to some varieties of English, potentially increasing the likelihood of Dutch speakers employing other cues in IDS to express their emotion. Moreover, Dutch IDS displays more *F0* variability across utterances compared to ADS [11], suggesting that increased variability may be observed in other cues as well.

An important consideration in the study of voice quality concerns its acoustic measurement. Voice quality can be assessed from acoustic data by calculating the amplitude difference between the first harmonic and the second harmonic in the spectrum (H1-H2) [24, 25]. High H1-H2 values indicate more breathy phonation and low values indicate more creaky phonation. However, this H1-H2 measure can be affected by both *F0* and F1 [26]. Even the H1*-H2* measure, which has been proposed to reduce the effect of vocal tract characteristics on the amplitude of harmonics, shows sensitivity to and a complex relationship with both *F0* and F1 [27, 28]. It is conceivable that the assessment of voice quality from H1-H2 (or H1*-H2*) in IDS could be affected by the higher and more variable *F0* and F1 in this register (see above). However, the recent observation that H1-H2 is higher in IDS was made without taking *F0* and F1 into consideration [13]. It is thus an open question whether the higher H1-H2 in IDS is better understood as a side-effect of known IDS characteristics, or a true reflection of a voice quality change. Similarly, one could only validly conclude that voice quality is more variable in IDS compared to ADS, if the observed variability extends beyond the variability in *F0* and F1. The present study thus aims to assess whether Dutch IDS is more breathy (research question 1) and has a more variable voice quality (research question 2) beyond the *F0* and F1 modifications in this register.

2. Methods

2.1. Participants and procedure

The materials used in this study were the IDS and ADS recorded from 19 native Dutch mother-infant dyads (10 daughters, 9 sons; infant mean age: 421 days; range: 322-476 days). IDS was elicited during a 10-minute parent-child interaction, for which the parent was instructed to play with their infant by unpacking and naming three bags of toys. ADS was elicited during a 10-minute parent-experimenter conversation about the play session. Recordings took place in a soundproofed studio with an omni-directional head-mounted Samson QV microphone into Enosoft DV Processor, sampled at 44,000Hz. [11] reported utterance-level *F0* in this corpus.

2.2. Coding

Waveforms and spectrograms were inspected to mark the starts and ends of the 1643 tokens [i, u, a:, ɑ] in the toy names. Coders excluded 695 of these tokens from further analysis, for example due to sound overlap or an unusual voice quality.

2.3. Acoustic analysis

The H1-H2 acoustic measurements were conducted with a publicly available Praat script [29]. *F0* and the first three formants were measured at the midpoint of each vowel, and the *F0* (+/-10%) was used to find harmonics in LTAS. If *F0* and all three formants could be computed, the maximum amplitude within the frequency boundaries around H1 and H2 were queried and the uncorrected H1-H2 calculated.

2.4. Final corpus size

Another 7 tokens were excluded due to failed acoustic measurements. Outliers were identified in the remaining 941 segments for H1-H2, F1, and *F0* separately. Within individual speakers, outliers were tokens more extreme than 1.5*IQR (interquartile range) across the four vowels and two registers. Across speakers, but within each vowel and register, outliers were tokens extending beyond 1.5*IQR. The distribution of the 149 outliers is reported in the results. A total of 792 tokens remained for the statistical analyses (see Table 1 for details).

2.5. Statistical analysis

Mixed-effects regressions models were conducted using *R*'s *lme4* package [30, 31]. Random-effects structures consisted of by-subject random intercepts, selected as optimal following the procedures in [32]. The significance of fixed effects was assessed with the Satterthwaite approximation of degrees of freedom using the *lmertest* package [33]. One R-implemented *t*-test was conducted.

Table 1. *The total number of tokens (min-med-max across individual speakers) included in the analyses.*

	IDS	ADS
[i]	201 (1-5-17)	55 (1-2-7)
[u]	263 (1-7.5-20)	74 (1-2-12)
[a:]	205 (1-7-17)	62 (1-2.5-9)
[ɑ]	557 (3-10-33)	82 (1-3-6)

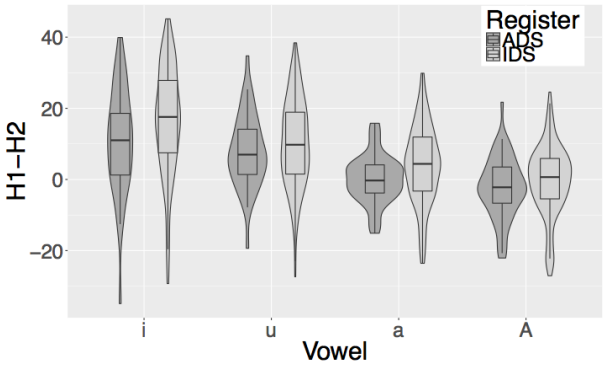


Figure 1. *Boxplots (whiskers extending to 1.5*IQR, violin outlines) of H1-H2 (not accounting for F0 and F1).*

3. Results

Research question 1, asking whether IDS is breathier than ADS, was first addressed without controlling for effects of *F0* and F1. The dependent variable in the analysis was H1-H2. Fixed factors were Register (contrast-coded: ADS=-1; IDS=1) and Vowel (deviation coded, with [ɑ] as the baseline that is not compared to the grand mean). The results revealed that Dutch IDS has a higher H1-H2 than ADS ($\beta_{Register}$=2.01, $t_{783.415}$=4.030, p<0.001). H1-H2 also differs across vowels, being high for high vowels [i] and [u] (β_i=7.573, $t_{781.307}$=8.312, p<0.001; β_u=3.142, $t_{780.660}$=3.621, p<0.001) and low for low vowel [a:] ($\beta_{a:}$=-3.557, $t_{784.000}$=-4.141, p <0.001). There was no evidence for a Register x Vowel interaction. Figure 1 illustrates that H1-H2 is higher in IDS compared to ADS in each vowel. These findings replicate those in [13].

A second analysis on H1-H2 with Register as fixed factor also included the 1st through to fourth-order polynomials of F1 and *F0* as continuous predictors that also interacted with Register. Critically, this analysis did *not* detect that H1-H2 was significantly higher in IDS compared to ADS ($\beta_{Register}$=0.354, $t_{757.934}$=0.747, p=0.455). These results suggest that an across-the-board increased H1-H2 in IDS could be (partially) a side-effect of the F0 and F1 characteristics instead of indexing increased breathiness.

In these analyses with F1 and *F0*, both were significant predictors of H1-H2 across four (F1) or three (F0) polynomials (main effects not reported). Figure 2 illustrates the (linear) relationship between *F0* and H1-H2. These results confirm the complex relationship of H1-H2 with F1 and *F0*

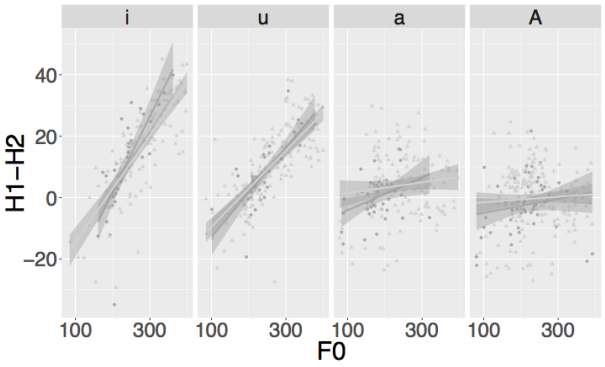


Figure 2. *Scatter plots illustrating the association between F0 and H1-H2 in ADS (dark) and IDS (light).*

Table 2. *Percentage of outlying tokens per acoustic measure across the 941 acoustically analysed tokens.*

	IDS	ADS
H1-H2	8.90	0.20
F1	6.30	3.36
F0	3.15	0.24

[28]. Moreover, these relationships were different across IDS and ADS (interactions not reported). To explore these interactions, the analysis of H1-H2, statistically accounting for F1 and *F0*, was conducted within each of the four vowels. The results revealed that [i] is breathier in IDS than ADS ($\beta_{Register}$=37.874, $t_{133.579}$=2.616, p=0.009). The effect of Register was marginally significant for [a:] ($\beta_{Register}$ =5.640, $t_{158.867}$=1.807, p =0.073) and not significant for [u] or /ɑ/ ($|\beta_{Register}|$<0.6, |t|<0.7). These results reveal that changes to voice quality in IDS may be vowel specific.

Research question 2 asked whether voice quality is more variable in IDS compared to ADS. The distribution of outliers (see Table 2 for details) provides initial evidence to this effect: IDS contained substantially more outliers than ADS, particularly on the H1-H2 measure. These outliers were equally divided over a high and low H1-H2. Increased voice quality variability in the IDS within the analysed section of the corpus is suggested by the elongated violin plots (Figure 1). However, neither of these observations account for variability within speakers, or the effects of F1 and *F0* on H1-H2.

To obtain a proxy to voice quality beyond F1 and *F0*, we extracted the residuals from a mixed effects model with H1-H2 as dependent variable and the first through to fourth-order polynomials of F1 and *F0* as predictors. A further 25 tokens were identified as outliers within these residuals and excluded, leaving 768 tokens for the final set of analyses. The within-speaker standard deviations of these H1-H2 residuals can be taken as a measure of voice quality variability within speakers and registers. Two speakers only contributed 0 or 1 token in ADS, making it impossible to compute standard deviations and leaving 17 speakers for the analysis of within-speaker variability. A paired-samples *t*-test on the standard deviations of the H1-H2 residuals revealed that mothers' voice quality was more variable in IDS compared to ADS (t_{16}=4.2845, p<0.001). Figure 3 displays these standard deviations, illustrating that IDS is spoken in a more variable voice quality, even after statistically accounting for variability due to F1 and *F0*, and that 15 of the 17 speakers followed this trend.

4. Discussion

This study set out to characterise the difference between Dutch infant-directed speech (IDS) and adult-directed speech (ADS) in terms of voice quality. Previous work on Japanese has suggested that IDS is breathier than ADS [13]. Although it appeared as if IDS is breathier in Dutch as well (research question 1), the observed H1-H2 increase may have been largely a side effect of *F0* and F1 modifications in this register. After statistically accounting for *F0* and F1, IDS was only significantly breathier in the vowel [i], marginally breathier in [a:], and not significantly breathier in [u] and [ɑ]. It is therefore currently unclear whether IDS is overall breathier than ADS. Despite the lack of an across-the-board difference, these results are the first to show that, even after statistically accounting for fluctuations in F0 and F1, voice quality is more variable in IDS compared to ADS (research question 2).

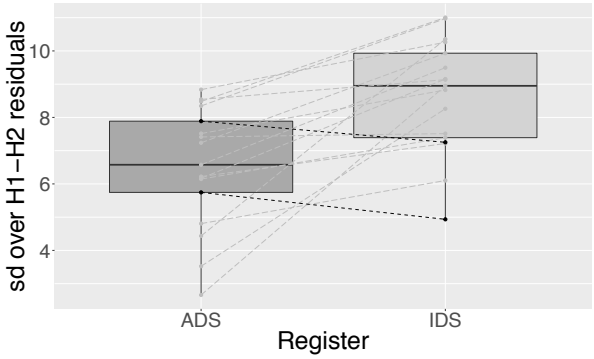


Figure 3. *Boxplots (whiskers extending to 1.5*IQR) of by-speaker standard deviations of H1-H2 residuals. Grey lines represent speakers with a higher sd in IDS, and black lines those with a higher sd in ADS.*

Beyond the cautionary note that the absence of evidence for an effect in frequentist statistics is not evidence for the absence of the effect, these findings present methodological considerations for future work on voice quality in IDS. Firstly, we excluded vowels with an "unusual voice quality", an umbrella category that included whispered speech. If very breathy tokens were inadvertently excluded as whispered, the present results may underestimate the breathiness of modal-sounding Dutch IDS. Conversely, [13] do not mention excluding whispered utterances from their analyses, even though those can make up 4.8%-18% of maternal IDS [34, 35]. The results in [13] may thus present an overestimation of the breathiness in modal-sounding Japanese IDS. Future work will need to carefully weigh the arguments for in- versus excluding whispered speech when characterising voice quality in IDS.

Second, this study presents a first step toward studying voice quality in IDS while incorporating recent insights about the relationship of H1-H2 with *F0* and F1 [28]. Our findings resonate the complexity of those relationships, and future voice quality research in IDS would benefit from a further improved understanding thereof. A partial solution to these measurement issues could be found in complementary methods of assessing voice quality, including (expert) perceptual ratings of perceived breathiness [36, 37] and production measures of the speakers' laryngeal activity [38].

The novel observation of variability in voice quality in IDS contributes to a larger literature on increased variability in IDS in other acoustic measures [5, 6, 7]. The increased variability in formant frequencies has been interpreted as contributing to the didactic properties of IDS [12]. However, voice quality does not serve a linguistic function in Dutch, and increased variability along this dimension suggests that parents (implicitly) engage in more vocal play when interacting with infants. Like *F0* variability, an unpredictable voice quality could evidence parents' high energy when speaking IDS [11, 15], and could render this register more interesting for the infant listener [11, 14].

To summarise, the present study has found that Dutch IDS may not be across-the-board breathier than ADS, but is spoken with more variability in its voice quality. This variability suggests that the "happy" sound of IDS may reflect high arousal to the same extent or more, than a positive valence. These results raise questions about the role that voice quality and variability play in the attuned parent-child interaction.

5. Acknowledgements

Corpus creation was supported by a Toptalent grant awarded to TB by the Netherlands Organization of Scientific Research. Thanks to Chad Vicenik for publicly sharing his Praat script [29] and to the Macquarie University Phonetics Lab, in particular Ivan Yuen, for helpful discussions during preparation of this manuscript.

6. References

[1] C.A. Ferguson, "Baby talk in six languages," *American anthropologist, (Special Publication)*, vol. 66, pp. 103–114, 1964.

[2] L.J. Trainor, C.M. Austin, and R.N. Desjardins, "Is infant-directed speech prosody a result of the vocal expression of emotion?," *Psychol Sci*, vol. 11, no. 3, pp. 188-195, 2000.

[3] R.P. Cooper and R.N. Aslin, "Preference for infant-directed speech in the first month after birth," *Child Dev.*, vol. 61, no. 5, pp. 1584-1595, 1990.

[4] L. Singh, J.L. Morgan, and C.T. Best, "Infants' listening preferences: Baby talk or happy talk?," *Infancy*, vol. 3, np. 3, pp. 365-394, 2002.

[5] M. Soderstrom, "Beyond babytalk: Re-evaluating the nature and content of speech input to preverbal infants," *Dev. Rev.*, vol. 27, no. 4, pp. 501-532, 2007.

[6] K.T. Englund and D.M. Behne, "Infant directed speech in natural interaction - Norwegian vowel quantity and quality," *J. Psycholinguist Res.*, vol. 34, no. 3, pp. 259–280, 2005.

[7] T. Benders, "Mommy is only happy! Dutch mothers' realisation of speech sounds in infant-directed speech expresses emotion, not didactic intent," *Infant Behav Dev*, vol. 36, no. 4, pp. 847-862, 2013.

[8] T. Banziger and K.R. Scherer, "The role of intonation in emotional expressions," *Speech Communication,* vol. 46, pp. 252–267, 2005.

[9] M. Goudbeek and K. Scherer, "Beyond arousal: Valence and potency/control cues in the vocal expression of emotion," *J. Acoust. Soc. Am.*, vol. 128, pp. 1322–1336, 2010.

[10] A. Warren-Leubecker and J.N. Bohannon III, "Intonation patterns in child-directed speech: Mother-father differences," *Child Dev.*, pp. 1379-1385, 1984.

[11] Anonymous, "Dutch fathers' infant-directed speech is characterised by highly variable and unpredictable pitch", manuscript in preparation.

[12] P.K. Kuhl, J.E. Andruski, I.A. Chistovich, and L.A. Chistovich, "Cross-language analysis of phonetic units in language addressed to infants" *Science Research Library*, vol. 277, pp. 684-686, 1997.

[13] K. Miyazawa, T. Shinya, A. Martin, H. Kikuchi, and R. Mazuka, "Vowels in infant-directed speech: More breathy and more variable, but not clearer," *Cognition*, vol. 166, pp. 84-93, 2017.

[14] O. Räsänen, S. Kakouros, and M. Soderstrom, "Connecting stimulus-driven attention to the properties of infant-directed speech—Is exaggerated intonation also more surprising?," *In Proc. 39th Ann. Conf. CogSci*, pp. 998-1003, 2017.

[15] S. Yildirim, M. Bulut, C. Lee, and A. Kazemzadeh, "An acoustic study of emotions expressed in speech," *Proc. InterSpeech*, pp. 2193–2196, 2004.

[16] E.A. Piazza, M.C. Iordan, and C. Lew-Williams, "Mothers Consistently Alter Their Unique Vocal Fingerprints When Communicating with Infants," *Current Biology*, vol. 27, pp. 3162-3167, 2017.

[17] J. Laver, *The Phonetic Description of Voice Quality*, Cambridge University Press, Cambridge, UK, 1980.

[18] M. Gordon and P. Ladefoged, "Phonation types: A cross-linguistic overview," *J Phon,* vol. 29, pp. 383–406, 2001.

[19] B.R. Gerratt and J. Kreiman, "Toward a taxonomy of nonmodal phonation," *J Phon,* vol. 29, pp. 365–381, 2001.

[20] 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.

[21] M. Ito, "Breathiness and politeness: An experimental study of male speakers of Japanese," *In 15th ICPhS,* pp. 2165-2168, 2003.

[22] C. Tsurutani and S. Shi, "How native speakers of Japanese try to sound polite," *JJL*, vol. 34, pp. 127-155, 2018.

[23] E.K. Johnson, M. Lahey, M. Ernestus, and A. Cutler, "A multimodal corpus of speech to infant and adult listeners," *J. Acoust. Soc. Am*, vol. 134, EL534-EL540, 2013.

[24] C.G. Henton and R.A.W Bladon, "Breathiness in normal female speech: Inefficiency versus desirability," *Language and Communication*, vol. 5, pp. 221–227, 1985.

[25] M. Garellek and P.A Keating, "The acoustic consequences of tone and phonation interactions in Jalapa Mazatec," *JIPA,* vol. 41, pp. 185–205, 2011.

[26] M. Epstein and B. Payri, "The effects of vowel quality and pitch on spectral and glottal flow measurements of the voice source," *J. Acoust. Soc. Am,* vol. 109, pp. 2413, 2001.

[27] M. Iseli, Y.L. Shue, and A. Alwan, "Age, sex, and vowel dependencies of acoustic measures related to the voice source," *J. Acoust. Soc. Am.,* vol. 121, pp. 2283-95, 2007.

[28] J. Kuang, "Covariation between voice quality and pitch: Revisiting the case of Mandarin creaky voice," *J. Acoust. Soc. Am,* vol. 142, pp. 1693, 2017.

[29] C. Vicenik. Los Angeles, CA, USA. Praat Voice Sauce Imitator [Online]. Available: http://phonetics.linguistics.ucla.edu/facilities/acoustic/PraatVoiceSauceImitator.txt

[30] D. Bates, M. Mächler, B. Bolker, and S. Walker, *Fitting linear mixed-effects models using lme4*, arXiv preprint arXiv:1406.5823, 2014.

[31] R Core Team, *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria, 2013.

[32] D. Bates, R. Kliegl, S. Vasishth, and H. Baayen, *Parsimonious mixed models*, arXiv preprint arXiv:1506.04967, 2015.

[33] A. Kuznetsova, P.B. Brockhoff, and R.H.B. Christensen, *Package 'lmerTest'*, R package version, 2(0), 2015.

[34] A. Fernald and T. Simon, "Expanded intonation contours in mothers' speech to newborns," *Dev. Psych*, vol. 20, pp. 104, 1984.

[35] M. Papousek, H. Papousek, and M.H. Bornstein, "The naturalistic vocal environment of young infants: On the significance of homogeneity and variability in parental speech," *Social perception in infants*, pp. 269-297, 1985.

[36] J. Hillenbrand, R.A. Cleveland, and R.L. Erickson, "Acoustic Correlates of Breathy Vocal Quality," *J Speech Lang Hear Res,* vol. 37, pp. 769-778, 1994.

[37] J. Kreiman and B.R. Gerratt, "Perceptual sensitivity to first harmonic amplitude in the voice source," *J. Acoust. Soc. Am*, vol. 128, pp. 2085–2089, 2010.

[38] C. Esposito, "An acoustic and electroglottographic study of White Hmong tone and phonation,", *J. Phonetics,* vol. 40, pp. 466-476, 2012.

Expression of affect in infant-directed speech to hearing and hearing impaired infants

Isabel Lopez and Christa Lam-Cassettari

Western Sydney University, MARCS Institute for Brain Behaviour and Development

c.lam-cassettari@westernsydney.edu.au

Abstract

Infant-directed speech (IDS) modifications are adversely affected by infant hearing loss [1, 2]. This study uses a 2-dimensional model [3] to rate the quality of emotion and arousal in mother's IDS to infants with normal hearing, simulated moderate hearing loss and simulated profound hearing loss. Results from a ratings study show that maternal IDS was rated more positive and arousing to infants with simulated profound hearing loss compared to normal hearing. These findings indicate that despite a reduction in vowel hyperarticulation to infants with simulated hearing loss, IDS exhibits increased positive affect as infant hearing loss is induced.

Index Terms: infant-directed speech, infant hearing loss, positive affect, arousal, mother-infant interaction

1. Introduction

Parents convey a range of emotions and intentions to their child during social interactions by adapting the quality of their voice. The distinctive sing song intonation we use to talk to babies is known as infant-directed speech (IDS) [4], and it is well established that infants prefer to listen to IDS over the more monotone style adult-directed speech (ADS) [5]. It is argued that IDS has three distinct functions: firstly, engaging infants in social interactions and maintaining infant attention, secondly, communicating affect, and finally facilitating language acquisition. Interestingly, the vocal exaggerations in IDS are often accompanied by exaggerated facial expressions which have also been shown to convey distinct emotional messages [6]. Understanding factors that influence the expression of emotion in IDS will provide new understanding of features that enhance the quality of early mother-infant communication.

The expression of vocal affect is an important component of social interactions. The stress that is expressed in the pitch and intensity of the talker's voice and the proportion of hesitation are considered indicators of the expression of affect [7]. High pitch and expanded pitch range are closely tied with the expression of positive emotions [8,9], and are used in IDS to encourage, reward and regulate the arousal level of their infants [10]. Infants also contribute to these interactions through active affective responsiveness. For instance, when Werker and McLeod [11] asked adults to rate infants' behaviour on Likert scales of affective responsiveness using videos of infants listening to IDS and ADS, the infants were rated more responsive, interested, and 'cute and cuddly' listening to IDS than ADS [11]. Infants discriminate between utterances conveying different affective intent types and respond appropriately when they are played IDS utterances that convey positive versus negative affect in their own, or another language [e.g., 12,13]. Furthermore, English-learning 5-month-olds show the appropriate affective responses to approving and disapproving IDS contours in English, German, Japanese and Italian [12]. Thus, infants are not simply passive recipients of mothers' affective messages, they reinforce mothers' affective behaviour with their own behaviours and this supports a positive feedback loop of communication [13]. While research has shown that infants can distinguish between comfort, attention, approval [14, 15] and disapproval [12] it is less well understood whether mothers produce these distinct affective intent types when their infant has a sensory impairment.

Not all elements of IDS are present in interactions with infants that have hearing loss. Studies have shown that exaggerated pitch is expressed in IDS regardless of an infant's hearing ability. Conversely, vowel hyperarticulation (which has been shown to support speech perception) is only present when an infant has normal hearing and not present if the infant is hearing impaired [1, 2]. In fact, it appears that the degree of vowel hyperarticulation decreases as a function of infant hearing ability which may be a direct result of reduced infant responsiveness when hearing loss is induced [2]. Consequently, the authors argued that hearing impaired infants may be disadvantaged during early communications because they are not being exposed to the complete range of IDS modifications.

Mothers typically use the highly animated IDS register to communicate with infants. Given that infants are interested in positive affect IDS over ADS [16], testing the quality of IDS to infants with reduced hearing ability vs. normal hearing, offers an interesting comparison. More specifically, investigating the degree of reduced infant responsiveness and increased levels of caregiving required on the mother's part in attempt to successfully engage with infants with simulated hearing loss. When an infant has normal hearing, a mother's voice is a powerful medium to elicit attention and engage in social interactions necessary for scaffolding early infant development. It is crucial to examine the quality of mother-infant interactions as it tends to differ as a function of infant hearing ability.

It was argued that vowel hyperarticulation decreases when infant hearing loss is induced due to reduced infant responsiveness [2]. Thus, understanding how non-verbal communicative gestures may differ as a function of simulated hearing loss, is important. Non-verbal communicative behaviours provide information beyond the voice which contributes to the success of social interactions. There is some consensus that interactions with later diagnosed hearing impaired infants are less successful than those of their hearing peers. In interactions where infants hear normally, there is a tendency for infants to be more engaged, responsive [17] and display more signalling behaviours (smiling, greeting and reaching) compared to infants with hearing loss [18]. Due to limited responsiveness from a hearing impaired infant, it can create stressful situations where mothers tend to use less comforting behaviours and encourage more communication breakdowns compared to infants that can hear normally [17, 18]. In fact, mothers tend to be more dominant, intrusive and less encouraging with their hearing impaired infant compared

to infants that can hear normally, as these interactions are challenging. As a consequence, mothers tend to overcompensate by over-structuring infant's play [17, 19] and in turn, breaking that crucial positive feedback loop that supports successful dynamic interaction. Systematic evidence on the quality of non-verbal communicative behaviours when an infant has hearing loss is necessary. Understanding whether deficits in affective intent and arousal are present during interactions with hearing impaired infants will contribute to our understanding of the role of non-verbal communicative gestures in early mother-infant communication.

This study explores whether the expression of affect in mother's IDS differs to infants normal hearing and simulated hearing loss. In the current study, samples of mother's IDS taken from the Lam and Kitamura [2] database were continuously rated using a 2-dimensional emotional space (2DES). The 2DES quantifies the level of valence (affect) and arousal expressed in the prosody of IDS [3, 20].

Lam and Kitamura [2] found that mothers hyperarticulated their vowels to the greatest extent in the full audibility condition, vowel hyperarticulation was reduced in moderate audibility condition and no hyperarticulation was present in the inaudible condition. It is expected that listeners will perceive differences in the quality of affect and arousal expressed in the mother's voice according to the infant's hearing ability. It is hypothesised that mother's speech along with mother's and infant's non-verbal communicative behaviours will be rated as more positive and arousing to normal and simulated moderate hearing loss infants compared to infants with simulated profound hearing loss.

2. Method

2.1. Experiment 1

2.1.1. Participants

The current sample consisted of 43 participants, with a mean age of 25 years (range: 18-51 years; 33 female, 10 male), recruited from Western Sydney University, Bankstown Campus. Undergraduate Psychology students received course credit toward their 1st year Psychology subject in return for their participation. Additional participants were recruited on the University campus and paid thirty dollars for their participation. Participants had an average of 3 years' experience caring for children < 5 years (range experience: 0-18 years). All participants had normal hearing and spoke English.

2.1.2. Stimulus Materials

After excluding the first minute of all recordings to allow the mother to warm up to the recording environment, the subsequent 25 seconds of clear speech was extracted for all 48 mother-infant pairs recorded in Lam and Kitamura [2]. Background noise or non-speech sounds (e.g., coughs, microphone bumped, clicks, clapping, heavy breathing) were removed. Long pauses were reduced to 1.5 seconds to remove any unnatural silences from the speech samples. To ensure the segmental content of the speech samples would not influence participant ratings, speech recordings were low-pass filtered at 500 Hertz in Praat [21].

2.1.3. Ratings procedure and apparatus

Participants completed ratings using MARCS Institute software to record continuous and fixed interval responses on a 2 dimensional emotional space (2DES) that measures valence (affect) and arousal. Given that participants will complete Likert scales in Experiment 2 after watching mother and infant behaviours, by including Likert scales in Experiment 1 it will aid in checking reliability across ratings. Participants listened to low-pass filtered IDS samples through headphones and made responses on a laptop computer using mouse clicks. Participants completed four practice trials before continuing to 48 test trials. The talkers heard in the trial samples were not played again. Once a speech sample was played, it was not repeated. Participants were instructed to move their mouse cursor around the x-y axes for the duration of the trial (~25 seconds) and rate samples according to their perceptions of the level of i) *Valence* on the x-axis (from low = *sadness* to high = *happiness*), and ii) *Arousal* on the y-axis (from low = *calmness* to high = *excitement*) conveyed in each sample. On a new screen, participants were also instructed to complete two Likert scales by rating i) *Emotional Arousal/animation* from *1 (low)* = calmness to 7 (*high)* = excitement and ii) *Valence* from *1 (low)*= sadness to 7 (*high)* = happiness. The 2DES software automatically randomises stimulus presentation for each participant. The task took approximately 40 minutes.

2.2. Experiment 2

2.2.1. Participants

Forty adults were recruited from Western Sydney University to rate the non-verbal communicative behaviours of mothers and infants in [2]. Their mean age was 24 years (range: 18-51 years; 34 female, 6 male). Participants had an average of 3-4 years' experience caring for children < 5 years (range of experience: 0-18 years).

2.2.2. Stimulus materials

Stimuli comprised 48 video excerpts of mothers and infants (N = 96) corresponding to the excerpts taken for speech ratings in Exp 1. To ensure that participants focused exclusively on the non-verbal communicative behaviours (and not speech) depicted in the video recordings, all ratings were conducted with the audio turned off. To maintain natural interactive motion in the videos, each video excerpt was cut as close as possible to its matching audio counterpart in Exp 1, in Adobe Premiere, and were approximately 20-25 seconds in duration. The videos were divided into two sets containing mother and infant interactive behaviours for counterbalancing purposes (i.e., to ensure that babies are not compared to mothers animation levels). Half of the participants rated the mother videos first (infant videos second), and the other half rated the infant videos first (mother videos second).

2.2.3. Ratings procedure and apparatus

Participants used DMDX software with a purpose written script to record ratings of mother's and infant's non-verbal communicative behaviours, and present excerpts in random order [22]. Participants completed a practice trial before continuing to the 96 test trials; once a sample was played, it was not repeated. Each participant watched a 20-25 second silent video of a mother/ infant interaction and instructed to complete two Likert rating scales adapted from the 2DES ratings program in Experiment 1; that is, i) *Negative/ positive emotion* (valence) from -3 *(high negative) = sadness* to 0 = *neutral* to +3 *(high positive)= happiness*, and ii) *Arousal/ animation* from 1 (low) = *calmness* to 7 (high)= *excitement*, each of which appeared on a new screen. The task took approximately 70 minutes.

3. Results

ANOVAs were conducted to determine whether there were any differences in the quality of IDS and non-verbal communicative behaviours as a function of infant hearing ability. ANOVAs included two planned comparisons: the first compared the normal hearing and moderate hearing loss conditions, and the second compared the normal hearing to profound hearing loss condition. For all comparisons, the alpha level was set at .05 and ANOVA test assumptions were satisfactorily met in all experiments, corrections were made where appropriate.

3.1. Experiment 1

3.1.1. Continuous (2DES) Ratings

ANOVAs showed a main effect for hearing condition for ratings of affect, $F(2, 84) = 11.56$, $p < .001$, $\eta_p^2 = .22$, and ratings of arousal $F(2, 84) = 13.69$, $p < .001$, $\eta_p^2 = .25$, indicating that adults perceived differences in the level of affect and arousal according to the infant's audibility condition. Planned comparisons revealed higher scores for the moderate compared to normal hearing conditions for ratings of affect $F(1, 42) = 16.61$, $p < .001$, $\eta_p^2 = .28$, and arousal $F(1, 42) = 15.52$, $p < .001$, $\eta_p^2 = .27$; this was also the case for the profound simulation compared to normal hearing condition for ratings of affect $F(1, 42) = 18.23$, $p < .001$, $\eta_p^2 = .30$ and arousal $F(1, 42) = 26.24$, $p < .001$, $\eta_p^2 = .39$. Overall, the analyses show that mother's speech was rated more affective and arousing when infants only had access to IDS with reduced audibility and were in the moderate or profound hearing loss conditions (Figure 1).

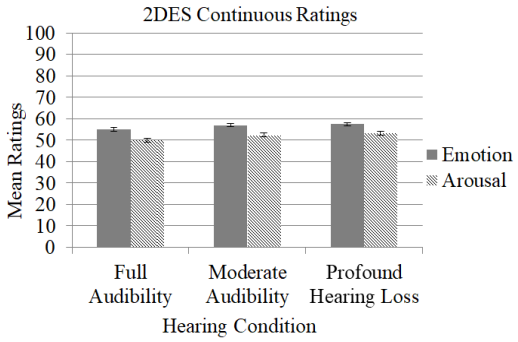


Figure 1: *Mean ratings of emotion and arousal in mother's IDS using the continuous 2DES model Error bars indicate SEM*

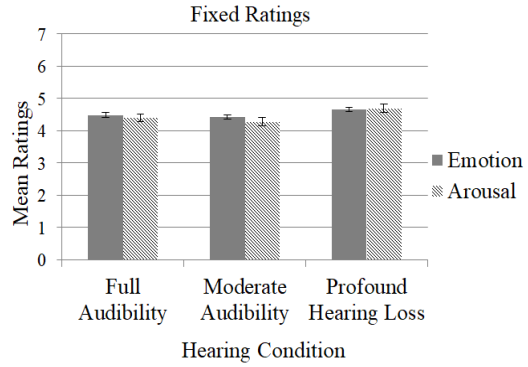


Figure 2: *Mean ratings of emotion and arousal in mother's IDS fixed ratings. Error bars indicate SEM*

3.1.2 Fixed Ratings

The ANOVAs showed a main effect for hearing condition in ratings of emotion $F(2, 84) = 6.08$, $p = .003$, $\eta_p^2 = .13$, and arousal $F(2, 84) = 17.26$, $p < .001$, $\eta_p^2 = .29$, indicating that fixed ratings of affect and arousal differed as a function of the infant's hearing condition. The planned comparisons between infants in moderate hearing compared to normal hearing condition showed greater affective ratings $F(1, 42) = .64$, $p = .428$, $\eta_p^2 = .02$, and lower arousal ratings $F(1, 42) = 2.18$, $p = .147$, $\eta_p^2 = .05$. The comparison between profound hearing loss and normal hearing condition showed larger ratings of affect $F(1, 42) = 10.85$, $p = .002$, $\eta_p^2 = .21$ and arousal $F(1, 42) = 15.25$, $p < .001$, $\eta_p^2 = .27$. Overall, the ratings indicate that mother's speech was more positive when infants had reduced audibility of the speech signal and were in the moderate or profound hearing loss conditions. This contrasted arousal ratings where mother's speech was more arousing when infants were in the normal hearing and profound hearing loss conditions (Figure 2).

3.2. Experiment 2

3.2.1. Mother's Non-Verbal Communicative Behaviours

Both ANOVAs showed a main effect for hearing condition in affect $F(2, 78) = 19.76$, $p < .001$, $\eta_p^2 = .35$, and arousal ratings $F(2, 78) = 5.86$, $p = .004$, $\eta_p^2 = .13$. The planned comparisons between infants in moderate hearing compared to normal hearing condition showed lower affective ratings $F(1, 39) = 16.53$, $p < .001$, $\eta_p^2 = .30$, and arousal ratings $F(1, 39) = 7.67$, $p = .009$, $\eta_p^2 = .16$; and lower in the profound hearing loss compared to normal hearing condition for affective ratings $F(1, 39) = 52.67$, $p < .001$, $\eta_p^2 = .58$ and arousal ratings $F(1, 39) = 9.88$, $p = .003$, $\eta_p^2 = .20$. Overall, the results revealed the level of affect and arousal in mothers' non-verbal communicative behaviours was greater when infants had full access to their mother's speech in the normal hearing condition, in comparison to the moderate and profound hearing loss conditions (Figure 3).

3.2.2. Infant's Non-Verbal Communicative Behaviours

Mauchley's test of sphericity yielded significant results ($p < .001$), in ratings of arousal in infant's non-verbal communicative behaviours, therefore the Huynh- Feldt adjustment was used as the epsilon value was above .75. The ANOVAs showed a main effect for hearing condition in affect $F(2, 78) = 117.99$, $p < .001$, $\eta_p^2 = .75$, and arousal ratings $F(1.40, 54.70) = 9.31$, $p = .001$, $\eta_p^2 = .19$, indicating infant's non-verbal ratings of affect and arousal differed with infant hearing conditions. The planned comparisons between infants in moderate and normal hearing condition showed lower affective ratings $F\ F(1, 39) = 5.96$, $p = .019$, $\eta_p^2 = .13$, and arousal ratings $F(1, 39) = 25.75$, $p < .001$, $\eta_p^2 = .40$; and lower in the profound hearing loss compared to normal hearing condition for affective ratings $F(1, 39) = 173.34$, $p < .001$, $\eta_p^2 = .82$ and arousal ratings $F(1, 39) = 3.42$, $p = .072$, $\eta_p^2 = .08$. These results showed that infants' non–verbal communicative behaviours were rated as more affective and arousing when infants were in the normal hearing condition compared to moderate and profound hearing loss conditions (Figure 3).

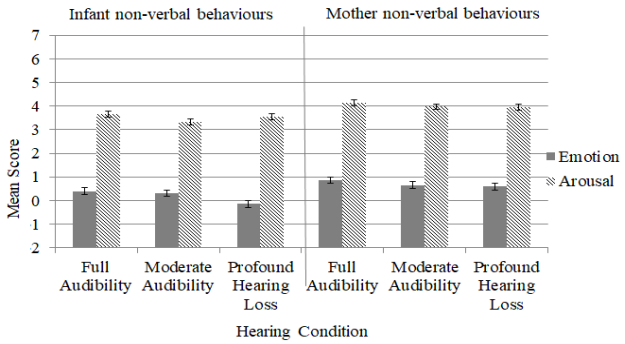


Figure 3: *Mean ratings of emotion and arousal in mother's and infant's non-verbal communicative behaviours in fixed ratings. Error bars indicate SEM.*

4. Discussion

This study showed that adults perceive differences in the prosodic quality of mother's IDS when an infant has simulated hearing loss. Positive affect and arousal increased as hearing loss was induced. This concurs with the findings of Lam and Kitamura [2] who showed that mothers continue to exaggerate fundamental frequency in IDS to infants with moderate and profound simulated hearing loss, despite a significant decline in vowel hyperarticulation as hearing loss was induced. It may be the case that as infants were not able to hear mother's IDS in the profound hearing loss simulation, they were less responsive and less involving with their mothers [11]. It is plausible that mothers tried to remedy the effects of reduced audibility and maintain the attention and responsivity of their infants by conveying more positive emotion and higher arousal in their intonation, and sacrificed vowel hyperarticulation in the process.

Looking at non-verbal communicative behaviours, the pattern of exaggeration concurs with previous evidence [2] that mothers and infants are more responsive when they have normal hearing. Mothers and infants were rated as expressing higher positive affect and arousal in their non-verbal communicative behaviours when infants were in the normal hearing versus simulated hearing loss conditions. In the normal hearing condition, infants may have been more responsive and highly engaged in mothers' interactive behaviours compared to conditions where hearing loss was induced. As such, infant responsiveness played a critical role in reinforcing mothers and encouraged mothers to vary verbal and non-verbal behaviours to different levels.

The presence of high positive affect and arousal may have also facilitated the communication of emotional messages and intent to the infant through the modification of facial expressions, voice and interactive behaviours [6] typical of IDS to normal hearing infants. As hearing ability is reduced, hearing impaired infants increasingly miss out on vital benefits of IDS such as communication of mother's motives, intention and current states. As such, the likelihood of communication breakdowns increases, making interactions more challenging. This study was limited to families with intact hearing undergoing a simulation of hearing loss, thus although results show similar trends to [1], they are not nationally representative and cannot be directly generalised to infants with hearing loss.

In summary, the quality of mother's speech as well as mother's and infant's non-verbal communicative behaviours differ as infant hearing loss is induced. When infants can hear normally, mothers and infants appear to convey higher positive affect and arousal in their non-verbal communicative behaviours which may be attributed to infant's responsiveness reinforcing mother's behaviours. Additionally, even though mothers make comparable modifications to the fundamental frequency and duration of IDS whether or not their infant can hear them [1, 2]; as infants lose the ability to hear and mothers reduce their vowel hyperarticulation, mothers appear to convey heightened positive affect and arousal in the tone of their voice. Indeed, it is plausible that as mothers are trying to maintain the attention of and infant that cannot hear properly and is less responsive, mothers unconsciously increase the expression of positive affect and arousal to repair the interaction and re-engage the infant. The current study provides a step forward in alleviating some of these negative effects on child development.

5. References

[1] Lam, C., & Kitamura, C. (2010). Maternal interactions with a hearing and hearing-impaired twin: Similarities and differences in speech input, interaction quality, and word production. *JSLHR., 53*(3), 543-555.

[2] Lam, C., & Kitamura, C. (2012). Mommy, speak clearly: Induced hearing loss shapes vowel hyperarticulation. *Dev. Sci., 15*(2), 212-221.

[3] Russell, J. A. (1980). A circumplex model of affect. *J. Pers. Soc. Psychol., 39*(6), 1161-1178.

[4] Fernald, A., Taeschner, T., Dunn, J., Papousek, M., 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.

[5] Cooper,R.P., & Aslin, R.N.(1990). Preference for infant-directed speech in the first month after birth. *Child Dev, 61*, 1584-1595.

[6] Chong, S. C. F., Werker, J. F., Russell, J. A., & Carroll, J. M. (2003). Three facial expressions mother direct to their infants. *Infant Child Dev., 12*(3), 211-232.

[7] Cowie, R., & Cornelius, R. R. (2003). Describing the emotional states that are expressed in speech. *Speech Com., 40*(1-2), 5-32.

[8] Scherer, K., R. (1986). Vocal affect expression: A review and model for future research. *Psych. Bull., 99*, 143-165.

[9] Trainor, L. J., Austin, C. M., & Desjardins, R. N. (2000). Is infant-directed speech prosody a result of the vocal expression of emotion? *Psych. Sci, 11*, 188-195.

[10] 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.

[11] 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. *Can.J.Psych,43*(2), 230.

[12] Fernald, A. (1993). Approval and disapproval: Infant responsiveness to vocal affect in familiar and unfamiliar languages. *Child Dev., 64*, 657-674.

[13] Kitamura, C., & Lam, C. (2009). Age Specific Preferences for Infant-Directed Affective Intent. *Infancy*, *14*(1), 77-100.

[14] 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 Dev., 67*(1), 205-217.

[15] Kitamura, C., & Burnham, D. (1998). The infant's response to maternal vocal affect. *Adv. Infancy Res., 12*, 221-236.

[16] Singh, L., Morgan, J. L., & Best, C. T. (2002). Infants' listening preferences: Baby talk or happy talk? *Infancy, 3*(3), 365-394.

[17] Wedell-Monnig, J., & Lumley, J. M. (1980). Child deafness and mother-child interaction. *Child Dev., 51*(3), 766-774.

[18] Koester, L. S. (1995). Face-to-face interactions between hearing mothers and their deaf or hearing infants. *Inf Beh.Dev,18*(2), 145-153.

[19] Brinich, P. M. (1980). Childhood deafness and maternal control. *J. Com. Dis., 13*(1), 75-81.

[20] Cowie, R., & Cornelius, R. R. (2003). Describing the emotional states that are expressed in speech. *Speech Com., 40*(1-2), 5-32

[21] Boersma, P., & Weenink, D. (2013). *Praat: Doing phonetics by computer.* Version 5.3.56, retrieved 19/6/13 http://www.praat.org

[22] Forster, K. I., & Forster, J. C. (2003). DMDX: A Windows display program with millisecond accuracy. *Beh.Res. Methods Instrum. Comput , 35*(1), 116-124.

F0 peaks are a necessary condition for German infants' perception of stress in metrical segmentation

Katharina Zahner & Bettina Braun

Department of Linguistics, University of Konstanz
{katharina.zahner|bettina.braun}@uni-konstanz.de

Abstract

Infants exposed to stress-timed languages have been demonstrated to use the stressed syllable to localize word beginnings. More recently, intonation has been shown to interfere with stress-based segmentation [1, 2]: For instance, stress-based segmentation was limited to accented words with bell-shaped accents that had the f0 peak aligned with the stressed syllable ("medial-peak" accents); segmentation failed when the f0 peak preceded or followed the stressed syllable in the target word's contour. Here, we test whether metrical segmentation is caused by the f0 peak on the stressed syllable or by the tonal alternation (LHL, bell-shaped contour in [2]). This allows us to probe whether f0 peaks are necessary cues for metrical segmentation. To this end, we replicated Zahner, et al. [2] but used cup-shaped intonation contours on the targets resulting in a tonal alternation in the opposite direction (HLH). Looking times obtained in a head-turn preference experiment showed no evidence of segmentation for the cup-shaped contours. This suggests that an f0 peak is a necessary condition for the stressed syllables to be used in stress-based segmentation, at least for German infants.

Index Terms: stress, pitch accent type, infant, German

1. Introduction

In order to extract units from speech, infants use both general strategies, e.g., transitional probabilities between syllables, and language-specific cues ([3, 4] for overviews). Regarding language-specific cues, prosodic properties of the ambient language shape segmentation behavior. For example, infants from stress-timed languages develop a stress-based strategy and interpret stressed syllables as word onsets, for Dutch [5-8], English [9-13], and German [14-16].

Two recent studies on German showed that intonation affects infants' segmentation behavior [1, 2]. Specifically, using electrophysiological measures, Männel and Friederici [1] demonstrated that word-form recognition is influenced by accentuation, i.e., whether or not a trochee (*Sirup* ['zi:.rʊp] 'syrup') receives a pitch accent. Infants' event-related potentials (ERPs) showed that the recognition of trochees differs with age and is modulated by accentuation: 6-month-olds only recognized trochees that were accentuated in familiarization, surfacing in a positive ERP response 500ms after word onset, i.e., before the end of the trochee. At 9 months, recognition was independent of accentuation and manifested in a (mature) negative response 400ms after word onset; this effect was followed by a late negativity only for accentuated words. At 12 months, infants recognized words independent of accentuation during familiarization (negative response 350ms after word onset). Hence, from that study, we may infer that accentuation generally facilitates segmentation.

Yet, a behavioural study by Zahner, et al. [2] suggests pitch accent type and the resulting consequences of (mis)alignment between the f0 peak and the stressed syllable (rather than accentedness in general) to modulate the segmentation success for German infants. Acoustically, pitch accent types differ in the alignment of the f0 peak in regard to the stressed syllable, making the position of f0 peak an unreliable cue to stress [17], see Fig. 1. In medial-peak accents (H*), used to introduce new information to the discourse [18], the f0 peak and the stressed syllable coincide. In early-peak accents (H+L*), signaling accessible information [19], the f0 peak precedes the stressed syllables, while it follows the stressed syllable in late-peak contours (L*+H / L* H-^H%), commonly employed for sentence-initial topics [20] or in (polar) questions [21].

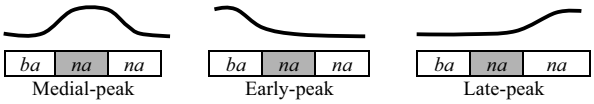


Figure 1: *Different pitch accent types on a trisyllabic word with stress on the second syllable (e.g., banana).*

Specifically, Zahner, et al. [2] familiarized German 9-months-olds with trisyllabic words (WSW stress pattern, e.g., *Lagune* [la.'gu:.nə] 'lagoon') in sentences; recognition of the embedded SW units (e.g., *gune* ['gu:.nə]) was tested. The WSW words were presented in one of 3 naturally occurring intonation conditions (between-subjects): medial-peak condition (peak-stress alignment), early- or late-peak contours (peak-stress misalignment), Fig. 1. Infants recognized the SW units only when f0 peak and the stressed syllable coincided. The authors concluded that a high-pitched accent (bell-shaped with the f0 peak on the stressed syllable, i.e., including a rising-falling movement) is important for stress perception and consequently segmentation.

Here, we test whether an f0 peak on the stressed syllable is a necessary cue for German infants to use a stressed syllable for metrical segmentation or whether it is the tonal alternation (LHL) that rendered the syllable with alternating pitch (the one with the high tonal target) particularly salient. In other words, infants may be particularly sensitive to the stressed syllable when the neighboring syllables differ in f0, e.g., LHL or HLH, but less sensitive when there is little change, LLH or HLL, as in the two misalignment conditions in [2]. This is compatible with an early view on stress perception on the utterance level by Bolinger [22], arguing that it is a "wide departure from a contour" (in any direction) that makes a syllable stand out [22: 112]. It would also be compatible with

a recent proposal on prominence perception according to which unexpected f0 events lead to increased prominence [23, 24]. In the current study, we tested cup-shaped contours, which have the tonal alternation in the opposite direction (HLH). If the change in f0 level is the necessary cue to trigger stress-based segmentation, *gune* is expected to be extracted in the cup-shaped contour. If, on the other hand, the f0 peak is the necessary cue, *gune* is not expected to be extracted. Another test case would be monotonous contours without any f0 movement. Likely, they would be perceived as unaccented, discarding them for use in the current paradigm (see [1]).

2. Experiment

2.1. Methods

2.1.1. Participants

Eighteen full-term infants (more than 37 weeks of gestation) from monolingual German families who finished the familiarization and all 12 test trials were included in the analyses (9 female, average age: 0;9.2 range: 0;8.18-0;9.17). Twelve further infants were tested but not included in the analysis due to crying (5), not attending to blinking lights (5), fussiness (1), and interference of a sibling (1).

2.1.2. Materials

For familiarization, we used the 4 passages from [2], which consisted of 6 sentences each. Each sentence contained the target once in different positions. The 4 target words were: *Kanone* [kʰa.ˈnoː.nə] 'cannon', *Lagune* [la.ˈguː.nə] 'lagoon', *Kasino* [kʰa.ˈsiː.no] 'casino', *Tirade* [tʰi.ˈraː.də] 'tirade'; (1) is an exemplar passage (SW part in trisyllabic carrier in bold; italics denote other accented words in the sentence).

(1) Hier entstand eine La**gune**. Die La**gune** war traumhaft. Die blaue La**gune** zieht Leute an. Eine kleine La**gune** ist schön. Seine La**gune** lag im Süden. Sie fotografierte ihre La**gune**.

'Here originated a lagoon. The lagoon was wonderful. The blue lagoon attracts people. A small lagoon is nice. His lagoon was situated in the South. She took a photo of her lagoon.'

A female speaker, who was trained in intonational phonology, recorded the passages, realizing the WSW carrier words (e.g., *Lagune*) with a cup-shaped intonation contour (HL*H), Fig. 2. We used natural productions instead of resynthesized stimuli to ensure the same stimulus quality across experiments. Note that the (tritonal) contour is not described in the German intonational system [21], but it is indeed present in German infant-directed speech (henceforth IDS), occurring in 7% of the 426 accentual movements in the KIDS Corpus [25]. We matched the current WSW words closely to those in Zahner, et al. [2], see Tab. 1.

For test, the 4 lists from [2] were used. They consisted of 15 tokens of the SW part of the WSW word, i.e., 15 tokens of [ˈguːnə], 15 tokens of [ˈraːdə], 15 tokens of [ˈnoːnə], and 15 tokens of [ˈsiːno], respectively. The trochees were falling and had an inter-stimulus interval of 800ms, see [2: 1346].

2.1.3. Procedure

The procedure was identical to the one described in Zahner, et al. [2], which employed the classic head turn preference paradigm [26] in a sentence-word order [9]. Infants were tested in a three-sided black booth with blinking lights in the Baby Speech Lab at the University of Konstanz. Sound was infant controlled, such that a new trial was initiated when the infant looked away for more than 2 seconds (s) or at the end of the stimulus. An experimenter with tight-fitting headphones and masking music controlled the experiment from behind the booth. All parents gave written consent and filled in a questionnaire. During familiarization, infants listened to 2 (out of 4) passages until they had accumulated a listening time of at least 45 s to each passage (e.g., *Lagune*- and *Kasino*- passage). For test, all infants listened to the same 4 test lists (*gune*, *sino*, *rade*, *none*), 2 of which were the SW units of the familiarized words (thus familiar) and 2 of which were not (and thus novel to the infants). Familiarization was counterbalanced across participants such that half of the infants were familiarized with the *Lagune*- and *Kasino*-passages and the other half with the *Kanone*- and *Tirade*-passages. The pseudo-randomized order of presentation of stimuli (right, left) was identical to [2].

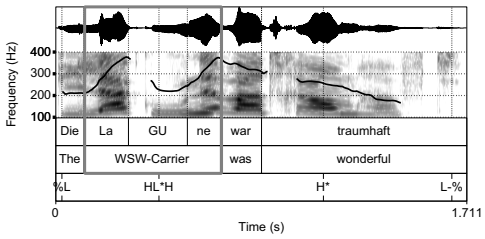


Figure 2: *Example target sentence in the cup-shaped intonation condition.*

Table 1. *Acoustic realization (mean values and standard deviations) of target words in the familiarization. Last column taken from [2: 1347].*

Acoustic variable	Cup-shape		Bell-shape [2]
F0 excursion of movement in st	HL* 8.4 (1.0)	L*H 8.5 (0.9)	LH* 8.8 (1.7)
Duration of first syllable (unstressed) in ms	196 (21)		182 (36)
Duration of second syllable (stressed) in ms	259 (20)		253 (25)
Duration of third syllable (unstressed) in ms	187 (50)		193 (66)
H1*-A3* ratio in middle of first vowel in dB	22.9 (7.4)		21.5 (6.5)
H1*-A3* ratio in middle of second vowel in dB	34.5 (10.7)		31.5 (14.1)
H1*-A3* ratio in middle of third vowel in dB	31.3 (12.4)		23.2 (5.3)

2.2. Results

Looking times were averaged by *familiarity status* (familiar vs. novel) for each infant. Infants looked 9.6s (sd = 3.5s) to familiar and 9.8s (sd = 3.4s) to novel test lists, see Fig. 3 (left). Nine infants out of 18 looked longer to the novel lists. A pairwise t-test indicated that the difference in looking times was not significant ($t(17) = 0.3$, $p > 0.78$). The null hypothesis was 3 times more likely than the alternative hypothesis (Bayes Factor = 0.3 [27]). To compare, infants in the bell-shaped condition in [2] looked 1.4s longer to novel lists, see Fig. 3 (right). To corroborate the difference across intonation

conditions, we pooled the data and tested for an interaction between *familiarity status* and *Experiment* in a repeated measures ANOVA with *familiarity status* as within-subjects factor and *Experiment* as between-subjects factor. The interaction between the two factors approached significance ($F(1,34) = 3.32$, $p = 0.08$). For the sake of completeness, Fig. 4 contrasts the 95% CIs of the looking time differences between the current study (left) and the study in [2] (right).

Hence, in contrast to the bell-shaped condition in [2], infants did not extract the trochaic part-word (SW) from the cup-shaped WSW carrier in this experiment. A mere alternation of tonal targets of opposite pitch height (HL*H) did not lead to a percept of lexical stress and the use in stress-based segmentation for the syllable that had the deviant pitch. The results obtained in the current experiment pattern with the misalignment conditions in Zahner, et al. [2], i.e., the early-peak and late-peak conditions in which the f0 peak is realized before or after the stressed syllable, respectively, speaking in favor of a high-pitched accented syllable that guides the segmentation process, at least in German infants.

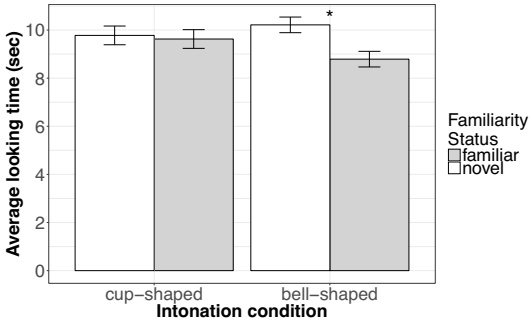


Figure 3: *Average looking times in the cup-shaped condition (left) and bell-shaped condition in [2] (right); whiskers represent ±1 standard error of the mean.*

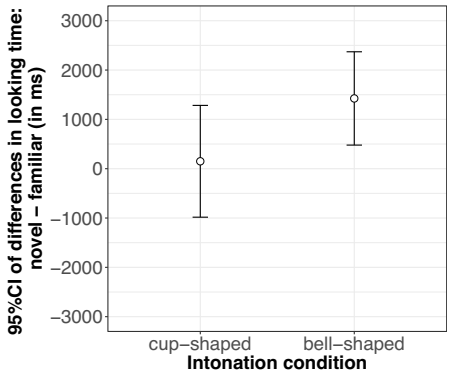


Figure 4: *Average looking time difference in the cup-shaped condition (left) and bell-shaped condition in [2] (right); whiskers represent the 95% confidence interval of the difference in looking time.*

3. General Discussion

The findings in Zahner, et al. [2] and in the current experiment together contribute to our knowledge on what leads to an infant's perception of a syllable as *stressed* [28, for discussion] and its use for segmentation. Zahner, et al. [2] showed that infants only succeed in extracting the trochee out of trisyllabic carrier words if the target is realized with a pitch accent type that renders the stressed syllable high-pitched, i.e., a medial-peak accent (bell-shaped, involving a rising-falling movement). Infants failed to extract the trochee when presented with early-peak or late-peak realizations on the target word. In the current experiment, infants also fail when the target is realized with a flipped medial-peak accent (cup-shaped) that pertains tonal alternation, but differs from the medial peaks in Zahner, et al. [2] in tonal height (LH*L vs. HL*H). These data suggest the f0 peak on a stressed syllable, comprising a rising-falling movement, to be a necessary cue for perceiving a syllable as stressed and thus as a word onset for German infants. The finding extends earlier research on accentedness and segmentation in German [1] by showing that high-pitched accent types are beneficial for segmentation.

Results from artificial language studies on linguistic grouping corroborate our explanation: [29, 30] show that infants tend to group syllable sequences that alternate in height (…H-L-H-L-H…) into trochaic units, taking the high-pitched syllables as the strong element. Currently, we see two explanations why high-pitched syllables (in our case high-pitched stressed syllables) are relevant for infants: first, medial-peak accents are most frequent in German IDS [25] (and also in German adult-directed speech [31]), which is why German infants most often encounter high-pitched stressed syllables, consequently fusing the concept of pitch peak and lexical stress; second, high-pitched stressed syllables might perceptually be more salient for German infants than low-pitched ones (which could either be a result of their acoustic nature or, rather circularly, a phenomenon which is prompted by the high frequency of occurrence itself). Teasing apart these two explanations is a challenge for future research.

Recent studies emphasized the relevance of lively and exaggerated contours in facilitating segmentation in German [32] and English [13]. These IDS productions show larger f0 excursions (phonetic modifications in pitch), along with durational differences. The findings in [2] and in the current paper demonstrate that different pitch accent types, i.e., alignment differences between peak and stress also modulate the segmentation success, with the f0 peak being a necessary cue to stress and segmentation. Future research needs to examine whether the segmentation success in the bell-shaped condition is also affected by the size of the f0 excursion.

We are currently investigating whether f0 peaks on unstressed syllables are sufficient to initiate the use of a syllable as a word onset in German infants, in addition to being a necessary cue. Visual-world eye tracking data from German (and Australian English) shows that this is indeed the case for adults: Adult listeners temporarily activated an initially stressed cohort competitor (e.g., *musical*) when a target, stressed on the second syllable (e.g., *museum*), was realized with an early-peak accent, i.e., an accent type in which the f0 peak preceded the stressed syllable [33, 34].

4. Acknowledgements

We thank Jana Neitsch, Stephanie Gustedt and Johanna Schnell for recording, recruiting and testing. We thank Janet Grijzenhout and Muna Schönhuber for discussion of the data.

5. References

[1] Männel, C. and Friederici, A. D., "Accentuate or repeat? Brain signatures of developmental periods in infant word recognition," *Cortex,* vol. 49, pp. 2788-2798, 2013.

[2] Zahner, K., Schönhuber, M., and Braun, B., "The limits of metrical segmentation: Intonation modulates infants' extraction

of embedded trochees," *Journal of Child Language,* vol. 43, pp. 1338-1364, 2016.

[3] Junge, C., "The proto-lexicon: Segmenting word-like units from the speech stream," in *Early word learning*, Westermann, G. and Mani, N., Eds., Abingdon, New York: Routledge, 2018, pp. 15-29.

[4] Saffran, J. R. and Kirkham, N. Z., "Infant statistical learning," *Annual Reviews,* vol. 69, pp. 181-203, 2018.

[5] Kuijpers, C. T., Coolen, R., Houston, D. M., and Cutler, A., "Using the head-turning technique to explore cross-linguistic performance differences," in *Advances in infancy research*. vol. 12, Rovee-Collier, C., Lipsitt, L., and Hayne, H., Eds., Stamford: Ablex, 1998, pp. 205-220.

[6] Kooijman, V., Hagoort, P., and Cutler, A., "Electrophysiological evidence for prelinguistic infants' word recognition in continuous speech," *Cognitive Brain Research,* vol. 24, pp. 109-116, 2005.

[7] Kooijman, V., Junge, C., Johnson, E. K., Hagoort, P., and Cutler, A., "Predictive brain signals of linguistic development," *Frontiers in Psychology,* vol. 4, p. 25, 2013.

[8] Junge, C., Kooijman, V., Hagoort, P., and Cutler, A., "Rapid recognition at 10 months as a predictor of language development," *Developmental Science,* vol. 15, pp. 463-473, 2012.

[9] Jusczyk, P. W., Houston, D. M., and Newsome, M., "The beginnings of word segmentation in English-learning infants," *Cognitive Psychology,* vol. 39, pp. 159-207, 1999.

[10] Cutler, A., Junge, C., Spokes, T., and Kidd, E., "Phonological acquisition: Stress-based segmentation in English," Paper presented at the Laboratory Phonology (LabPhon16), Lisbon, Portugal, 2018.

[11] Houston, D. M., Kuijpers, C., Coolen, R., Jusczyk, P. W., and Cutler, A., "Cross-language word segmentation by 9-month-olds," *Psychonomic Bulletin & Review,* vol. 7, pp. 504-509, 2000.

[12] Mason-Apps, E., Stojanovik, V., Houston-Price, C., and Buckley, S., "Longitudinal predictors of early language in infants with Down Syndrome: A preliminary study," *Research in Developmental Disabilities,* vol. 81, pp. 37-51, 2018.

[13] Floccia, C., Keren-Portnoy, T., DePaolis, R., Duffy, H., Delle Luche, C., Durrant, S., *et al.*, "British English infants segment words only with exaggerated infant-directed speech stimuli," *Cognition,* vol. 148, pp. 1-9, 2016.

[14] Höhle, B., "Der Einstieg in die Grammatik: Die Rolle der Phonologie/Syntax- Schnittstelle für Sprachverarbeitung und Spracherwerb [First steps into grammar: The role of phonology/syntax interface for language processing and acquisition]," Habilitationsschrift, Freie Universität Berlin, Berlin, Germany, 2002.

[15] Bartels, S., Darcy, I., and Höhle, B., "Schwa syllables facilitate word segmentation for 9-month-old German-learning infants," in *33rd Annual Boston University Conference on Language Development*, Somerville, M.A., 2009, pp. 73-84.

[16] Altvater-Mackensen, N. and Mani, N., "Word-form familiarity bootstraps infant speech segmentation," *Developmental Science,* vol. 16, pp. 980-990, 2013.

[17] Ladd, D. R., *Intonational phonology*, 2nd ed. Cambridge: Cambridge University Press, 2008.

[18] Kohler, K., "Terminal intonation patterns in single-accent utterances of German: Phonetics, phonology and semantics," *Arbeitsberichte des Instituts für Phonetik und Digitale Sprachverarbeitung der Universität Kiel (AIPUK),* vol. 25, pp. 115-185, 1991.

[19] Baumann, S. and Grice, M., "The intonation of accessibility," *Journal of Pragmatics,* vol. 38, pp. 1636-1657, 2006.

[20] Braun, B., "Phonetics and phonology of thematic contrast in German," *Language and Speech,* vol. 49, pp. 451-493, 2006.

[21] Grice, M., Baumann, S., and Benzmüller, R., "German intonation in autosegmental-metrical phonology," in *Prosodic typology. The phonology of intonation and phrasing*, Sun-Ah, J., Ed., Oxford: Oxford University Press, 2005, pp. 55-83.

[22] Bolinger, D., "A theory of pitch accent in English," *Word,* vol. 14, pp. 109-49, 1958.

[23] Kakouros, S., Salminen, N., and Räsänen, O., "Making predictable unpredictable with style – Behavioral and electrophysiological evidence for the critical role of prosodic expectations in the perception of prominence in speech," *Neuropsychologia,* vol. 109, pp. 181-199, 2018.

[24] Kakouros, S. and Räsänen, O., "Perception of sentence stress in speech correlates with the temporal unpredictability of prosodic features," *Cognitive Science,* vol. 40, pp. 1739-1774, 2016.

[25] Zahner, K., Schönhuber, M., Grijzenhout, J., and Braun, B., "Konstanz prosodically annotated infant-directed speech corpus (KIDS corpus)," in *Proceedings of the 8th International Conference on Speech Prosody*, Boston, USA, 2016, pp. 562-566.

[26] Kemler Nelson, D. G., Jusczyk, P. W., Mandel, D. R., Myers, J., Turk, A., and Gerken, L., "The head-turn preference procedure for testing auditory perception," *Infant Behavior and Development,* vol. 18, pp. 111-116, 1995.

[27] Wagenmakers, E.-J., "A practical solution to the pervasive problems of p values," *Psychonomic Bulletin & Review,* vol. 14, pp. 779-804, 2007.

[28] Bhatara, A., Boll-Avetisyan, N., Höhle, B., and Nazzi, T., "Early sensitivity and acquisition of prosodic patterns at the lexical level," in *The development of prosody in first language acquisition*. vol. 23, Esteve-Gibert, N. and Prieto, P., Eds., Amsterdam [u.a.]: John Benjamins, 2018, pp. 37-57.

[29] Bion, R. A. H., Benavides-Varela, S., and Nespor, M., "Acoustic markers of prominence influence infants' and adults' segmentation of speech sequences," *Language and Speech,* vol. 54, pp. 123-140, 2011.

[30] Abboub, N., Boll-Avetisyan, N., Bhatara, A., Höhle, B., and Nazzi, T., "An exploration of rhythmic grouping of speech sequences by French- and German-learning infants," *Frontiers in Human Neuroscience,* vol. 10, p. 292, 2016.

[31] Peters, B., Kohler, K., and Wesener, T., "Melodische Satzakzentmuster in prosodischen Phrasen deutscher Spontansprache - Statistische Verteilung und sprachliche Funktion [Meldodic sentence accent patterns in prosodic phrases of German sponteneous speech - Statistical distribution and linguistic function]," in *Prosodic structures in German spontaneous speech (AIPUK 35a)*, Kohler, K., Kleber, F., and Peters, B., Eds., Kiel: IPDS, 2005, pp. 185-201.

[32] Schreiner, M. S. and Mani, N., "Listen up! Developmental differences in the impact of IDS on speech segmentation," *Cognition,* vol. 160, pp. 98-102, 2017.

[33] Zahner, K., Kember, H., and Braun, B., "Mind the peak: When museum is temporarily understood as musical in Australian English," in *Proceedings of the 18th Annual Conference of the International Speech Communication Association (Interspeech)*, Stockholm, Sweden, 2017, pp. 1223-1227.

[34] Zahner, K., Schönhuber, M., Grijzenhout, J., and Braun, B., "Pitch accent type affects lexical activation in German: Evidence from eye tracking," in *Proceedings of the 16th Speech Science and Technology Conference (SST),* Sydney, Australia, 2016.

Estimation of vocal tract and trachea area functions from impedance spectra measured through the lips

Anne Rodriguez, Noel Hanna, André Almeida, John Smith and Joe Wolfe

School of Physics, University of New South Wales
n.hanna@unswalumni.com

Abstract

Determining the area function $A(x)$ of the airway between the lips and vocal folds from external measurements is a classic inverse problem. $A(x)$ is estimated by fitting the acoustic impedance measured through the lips. Excellent fits are possible with about eight cylindrical segments representing the tract. In examples where $A(x)$ has only small slope, moderately good agreement is found on the scale of about a centimetre. Calculations of the impedance loading the glottis are affected by the epilaryngeal tube, which has less effect on the impedance through the lips. The frequencies of these extrema are better estimated than their magnitudes.

Index Terms: vocal tract, acoustic impedance, speech production, subglottal tract

1. Introduction

Measurements of the acoustic impedance spectrum of the vocal tract can now be made through the lips rapidly, precisely and over a large frequency range. Best results are obtained with the glottis closed, but good results are possible during phonation and breathing. With no other information, how much can they tell us about the shape of the tract, the trachea and the acoustic load they impose on the glottal source?

The vocal tract is often analysed as an acoustic duct with varying cross section $A(x)$ along its length x – essentially a one dimensional (1D) model. Discrete $A(x)$ models have a long history. Fant [1] used a simple, two-cylinder model, to demonstrate the acoustic phonetic model of vowel placement: the position of the discontinuity between the two segments (of lengths l_1 and l_2) determined the frontness or backness of the vowel, while the area ratio (A_1/A_2) largely determined the vowel height. Sets of several or more cylinders approximating $A(x)$ are now commonly used, and used hereafter in this paper.

Determining the vocal tract $A(x)$ from the acoustic signal has long been a goal of speech science (e.g. [2,3]) but the speech signal, sparsely sampled in the frequency domain, does not contain enough information to do this reliably. Pulse reflectometry measures the impulse response of the tract at the lips while sealed around a measurement device; from this $A(x)$ can be reconstructed with several assumptions [4], this approach is used clinically.

The acoustic impedance is the (complex) ratio of acoustic pressure (p) to flow (U): $Z(f) = p/U$. Here we examine whether recent advances in impedance spectrometry are sufficient to estimate $A(x)$ reliably. Although the difficulty of the inversion problem is similar to pulse reflectometry, the advantage of this approach is the improved signal:noise ratio, which allows more detailed acoustic information to be used as an input.

In the present paper, we investigate solutions to the inversion problem in terms of the number of parameters that can reasonably be extracted from good data and their goodness of fit. It uses a known target $A_t(x)$ and a one-dimensional duct model to generate a target acoustic impedance spectrum $Z_t(f)$, which is then fitted by calculating $Z(f)$ from estimated functions $A(x)$ whose parameters are varied to minimise the squares of differences between $Z(f)$ and $Z_t(f)$. We concentrate on the vocal tract and use area functions $A_t(x)$ for the American English vowels /i/, /o/, /ɔ/, /æ/ and /ɛ/ from [5]. An intended application of the $A(x)$ is the back propagation calculation of the acoustic signal: obtaining the flow and pressure at the vocal folds, as described in a companion paper in these proceedings [6]. For this reason, the vocal tract shapes are modified to provide a consistent lip aperture.

Recent measurements of $Z_t(f)$ during a neutral vowel gesture show the effects of yielding, lossy walls [7], and measurements during inhalation show the effects of the subglottal duct (trachea) [8].

2. Materials and Methods

2.1. Acoustic impedance measurements

Measurements of $Z_t(f)$ through the lips are made as described in [7]. Briefly, the subject seals the lips around the end of a cylindrical waveguide on which are mounted three microphones, which are input via conditioning amplifiers and an audio interface to a computer.

At the opposite end is a loudspeaker in parallel with a short, narrow pipe that allows exhaled air to exit. The computer generates a sum of sine waves in the range 200 to 4000 Hz, with magnitudes and phases adjusted to improve signal:noise ratio, via an audio interface and amplifier; this drives the speaker. With suitable calibration, impedance at the lips may be calculated at each frequency from the three microphone signals.

This method measures the input impedance $Z_t(f)$ in a fraction of a second but, on its own, no corresponding $A(x)$.

2.2. Simulated impedance through the lips and glottis

Whether for measurements at the lips or a calculated load at the glottal source, impedance spectra for the $A(x)$ are calculated using the transfer matrix method [*e.g.* 9]. Starting with a load impedance Z_L, the impedance at the opposite end of the adjacent cylindrical section is calculated as

$$Z = \frac{\rho c}{\pi r^2}\left(\frac{\pi r^2 \cdot Z_L \cos(\Gamma l) + 2j\rho c \cdot \sin(\Gamma l)}{j\pi r^2 \cdot Z_L \sin(\Gamma l) + 2\rho c \cdot \cos(\Gamma l)}\right) \quad (1)$$

where ρ is the density of air, c the speed of sound, r the radius, l the length and Γ is the complex wave number

$$\Gamma = \frac{\omega}{c} - j\alpha \approx \frac{\omega}{c} - j\frac{(1.2\times10^{-5}\ s^{1/2})\sqrt{\omega}}{r} \quad (2)$$

where ω is the angular frequency and α is an attenuation coefficient [9], which accounts for losses to the tube walls by viscous drag and thermal conduction. For a closed, immoveable termination (a reasonable approximation to a closed glottis), the load impedance is infinite.

Figure 1 shows the radius $r(x)$, for several vowels from the published $A(x)$ in [5], assuming cylindrical segment geometry and modified to provide a consistent lip aperture.

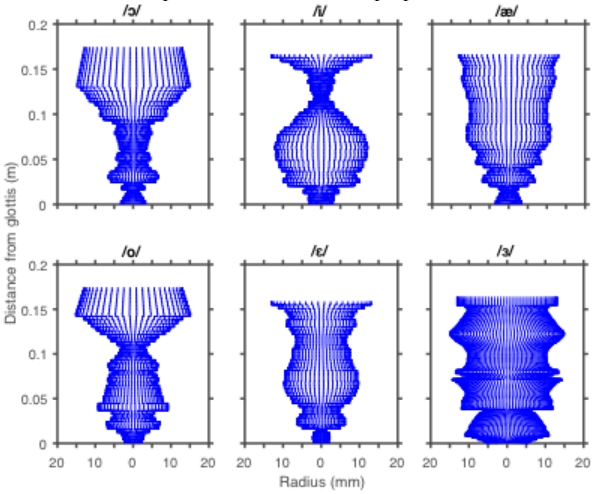


Figure 1. *Shows r(x) for the synthetic vowels /ɔ/, /i/, /æ/, /o/, and /ɛ/ modified from [5], and the vowel /ɜ/ from [10].*

For the case of the open glottis during inhalation, we use a very simple model for the trachea following [11]. Because the trachea branches rapidly, a single open cylinder with a large flange is used.

The $Z(f)$ that loads the glottal source was similarly calculated by iteration of the transfer matrix method, this time starting with the radiation impedance at the open lips and working back to the glottis.

2.3. Inversion and the optimal number of elements

We begin with an $A_t(x)$ (subscript t for target) from Figure 1 and calculate $Z_t(f)$. We then consider a candidate $A(x)$ function having n cylindrical elements with lengths and radii l_i and r_i, all of which may be adjusted during the fitting. From these, we calculate $Z(f)$ through the lips as described above. Scaled sums of squares of errors for magnitude, phase and combination respectively are calculated:

$$S_m = \frac{1}{N}\sum\left(\frac{\log(|Z|)-\log(|Z_t|)}{\log(|Z_t|)}\right)^2 \quad (3)$$

$$S_\phi = \frac{1}{N}\sum\left(\frac{\phi-\phi_t}{\pi}\right)^2 \quad (4)$$

$$S_\Sigma = \frac{1}{N}\sum\left(\frac{\log(|Z|)-\log(|Z_t|)}{\log(|Z_t|)}\right)^2 + K\left(\frac{\phi-\phi_t}{\pi}\right)^2 \quad (5)$$

where there are N frequencies. These are minimised using the MATLAB non-linear least squares fitting function *fit*. The range goes from 1 to 4 kHz for theoretical fits and targets, but over the measured range of 200-4000 Hz when fitting real data. The purpose of the semi-log scale for magnitude is to give impedance maxima and minima similar weight in the sum.

Minimising S_m or S_ϕ gives similar solutions for $A(x)$. We also minimised S_Σ. (The denominators in S_m and S_ϕ give two sums of order unity. K is a weighting factor set to 1/20, chosen to approximate the ratio S_m/S_ϕ in Figure 2). This gives an improvement to the A(x) fits; see Figure 2.

Inside a loop for the minimisation of S, l_i and r_i are varied within the constraints 0.1 to 100 mm and 0.1 to 30 mm respectively and with initial values from 5 to 50 mm and 10 mm.

3. Results and Discussion

3.1. Synthetic vocal tract models

3.1.1. Ideal number of cylindrical segments

The number n of cylindrical segments was varied from 1 to 10, and a minimum found for each n, so that S may be plotted as a function of n, as in Figure 2. This figure uses data from the vowel /ɔ/. For these data, and for the others studied, each successive added element makes an improvement up to about n = 7 or 8, but there is no improvement thereafter for S_m or S_ϕ. For the results that follow, n = 8.

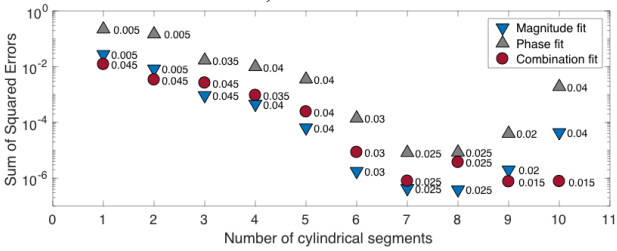


Figure 2. *Scaled sum of squares vs number of cylindrical segments n for fits for the vowel* /ɔ/, calculated at the lips. *Fits minimise differences in impedance magnitude, phase or combination of both. The value at each point shows the initial l_i that gave the lowest SSE.*

Figure 3a compares the radii $r_t(x)$ and the $r(x)$ of the n=8 model from Figure 2. On a longitudinal scale of about a centimetre, the geometrical features are roughly reproduced.

Fig 3b compares $Z_t(f)$ and $Z(f)$ through the lips. Here, the fit is encouragingly good – but this is expected for a fit whose difference has been minimised by adjusting 16 parameters.

The match between the load $Z_g(f)$ on the glottis calculated using $A(x)$ and $A_t(x)$ is not as good, of course: this quantity was not fitted. It is worth noting, however, that the frequencies of the peaks in Z_g, which are closely related to the formants outside the mouth, are fitted fairly well. One can understand some of the limitations in estimating the magnitude of impedance at the glottis Z_g: varying the cross section of a short constriction just above the glottis would be expected to vary a large inertance loading the glottis and in series with the tract: this would have a large effect on Z_g, but relatively little on the impedance at the lips. The fit is poorest at high frequency, in part because the 8-element fit has a lower spatial resolution than the 43-element target. It is also important to point out that, at frequencies above about 3 kHz, the approximation that wavelength >> radius, necessary for 1D plane wave propagation, is less reliable, so one might expect the 1D model to begin to fail.

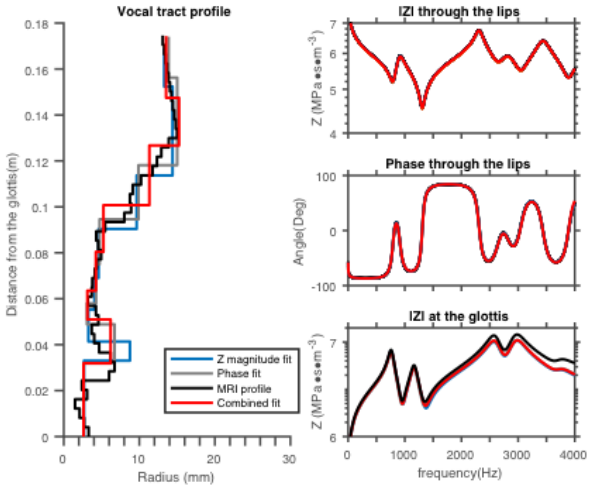


Figure 3. *(left) compares radii: target $r_t(x)$ (black) and fitted $r(x)$ for the vowel /ɔ/ having 43 and 8 elements respectively. (upper and middle right) compare the $Z(f)$ magnitude and phase through the lips calculated from these $r_t(x)$ and $r(x)$. (lower right) compares the calculated $Z_g(f)$ at the glottis.*

3.1.2. Sensitivity to initial parameters

Inversion problems in general have potentially many solutions. In this case, how much do they differ? To test this, we tried two methods. One was to vary the initial shape $A_0(x)$ (by changing l_i *or* r_i) and to compare the final $A(x)$. Increasing the initial l_i without constraining the sum of l_i improved the fit up to a certain point, then reduced the quality of fit. The results shown in Figure 2 show those whose initial lengths gave the best fit. Changing the initial r_i had little effect. Another was to compare the $A(x)$ determined by minimising S_m, S_ϕ and S_Σ. Although the sums of squares (quality of fit) varied, the $A(x)$ were qualitatively very similar in most cases (see Figures 4 and 5).

3.2. *In vivo* vocal tract measurements

The calculations above assumed rigid-walled ducts and losses appropriate to dry, hard walls. It is interesting to look at real measurements of $Z(f)$, which, at low frequencies, show features associated with walls of finite mass and rigidity. These require further parameters, for the specific mass, stiffness and losses in the wall. However, they also show new features: a new maximum and minimum in $Z_t(f)$, and so are a more demanding set to fit.

The data from one of the subjects in [7, 8] were used here. For this subject (S7), MRI measurements of the vocal tract and the upper section of the trachea were available for the sustained vowel /ɜ:/ [10], the neutral vowel in the word 'heard' in Australian English.

Figure 4 shows a measurement of $Z(f)$ through the lips with closed glottis. The shape is in qualitative (and even semi-quantitative) agreement with the impedance of a uniform 17 cm cylinder above about 300 Hz. However, it fails at low frequency, where a new maximum and minimum appear. The maximum at about 200 Hz is attributed to the mass of the tissue surrounding the vocal tract vibrating on the compliance ('spring') of the enclosed air. (An impedance minimum at about 20 Hz due to that of the same mass vibrating on the compliance of its own tissue is not visible in the measurement at this range [7].) Also superposed is a fit using the technique described above, but with three new parameters: the compact mass (inertance) and specific loss associated with its motion, and an attenuation coefficient (the α in equation (2)) which is increased to model the increased losses of wet, irregular wall losses in living tissue (discussed in [7]). These elements are added as a compact resonant circuit in parallel with the impedance calculated through the lips.

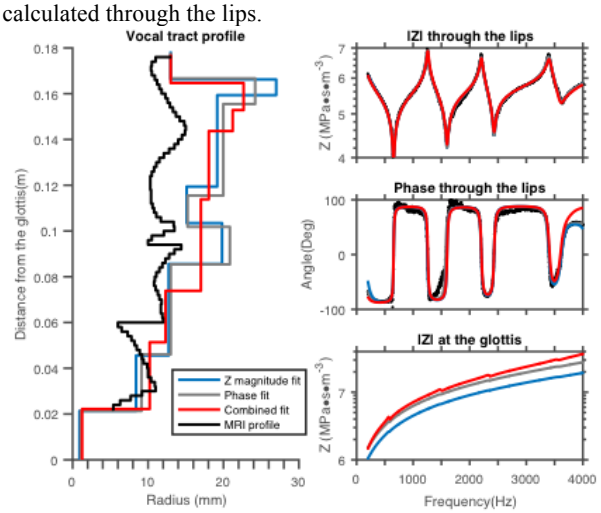


Figure 4. *Measured (black) and fitted impedance magnitude and phase spectra for a male subject with glottis closed. Fitted r(x) compared against r(x) from MRI derived geometry [10]. Otherwise, as Figure 3.*

Although this adds even more parameters, the effect of each of the parameters is readily apparent and almost independent.

- the mass adjusts the frequency of the first impedance peak
- the tissue loss adjusts the magnitude of the first impedance peak
- the wall loss adjusts the bandwidth of all extrema

The first element at the lip has a radius fixed at 13.1 mm (equal to that of the three-microphone impedance head used for measurements) and the upper bound for the length was decreased to 40 mm from 100 mm.

Note, if the very low frequency range below 50 Hz were to be considered, then an additional tissue compliance parameter would also be necessary for the fit.

Figure 4 shows the $r(x)$ based on the fits compared with $r(x)$ derived from MRI. Note that neither corresponds to the target $Z_t(f)$: the MRI is based on a 2D midsagittal image and so is also only an approximation of the true unknown target $A_t(x)$, since it makes the unrealistic assumption of 1D geometry. The major shape features are roughly reproduced, although the fitted radii are generally larger. The total length of the tract is different in the two cases, however, the 2 cm immediately above the glottis has little effect on the impedance through the lips. Its effect on the impedance seen by the glottis is discussed below.

Measurements made on a female subject performing a similar gesture are shown in Figure 5. Note the relatively high impedance of the first minimum compared to the second. This implies that the vocal tract is less well approximated by a uniform cylinder, and possibly that the resonance due to the total tract length terminated at the closed glottis is not as strong as that at the back of the mouth or pharynx, i.e. due to a constriction less sound power reaches and is reflected back from the glottis. Indeed, both fits to amplitude and phase give a large mouth cavity with a constricted pharynx.

Also note the effect of the epilaryngeal tube, the area just above the vocal folds that includes the aryepiglottic folds, on

the impedance at the glottis. The magnitude and phase fits both infer very narrow epilaryngeal tubes, which act as a large inertance on the glottal load (see also Figure 4). However, the combination fit produces a wider epilaryngeal tube, which has the effect of greatly increasing the relative prominence of the first three resonances. Note that here the true $A_t(x)$ is unknown so we cannot (yet) know which is better.

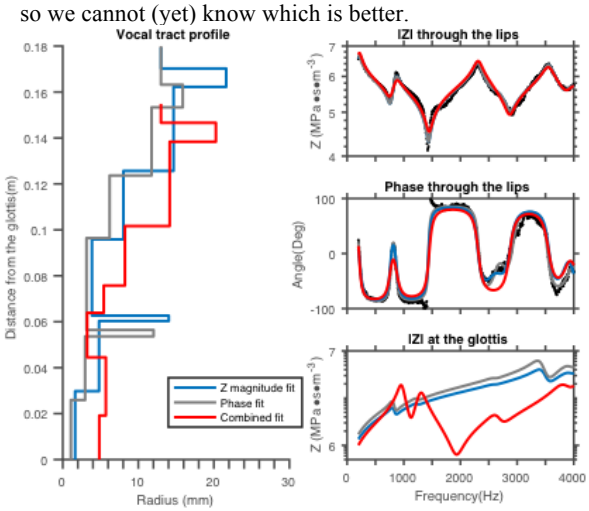


Figure 5. As Figure 4 *for a female subject with glottis closed. r(x) are estimated from fits, $r_t(x)$ is unknown.*

3.2.1. Including the subglottal tract

A preliminary trial of fitting to measurements with open glottis during inhalation, showed that at least three additional cylindrical elements are needed to represent the glottis and subglottal tract, which is terminated with an ideal flange, to represent the rapid branching as suggested by [11]. Without considering yielding walls, and using the same loss value as for the vocal tract, this adds six more parameters; three radii and three lengths. However, it gives several more features to fit: because it roughly doubles the length of the non-uniform vocal tract, the number of purely acoustic minima and maxima in $Z(f)$ is approximately doubled.

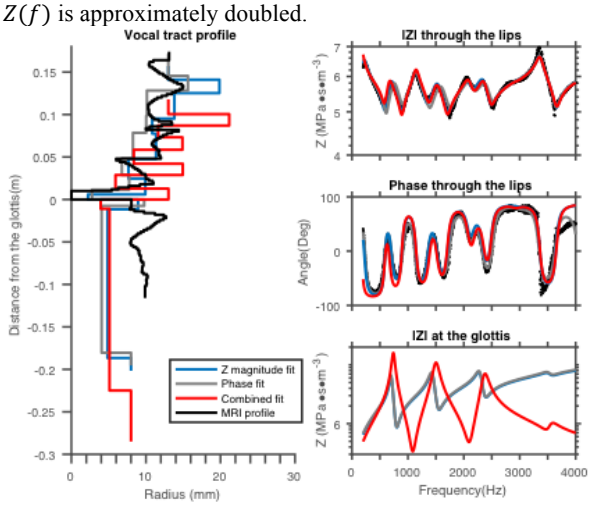


Figure 6. *As Figure 4 for a male subject during inhalation to allow estimation of the subglottal geometry. The end of the black MRI data is due to the MRI image ending about 12 cm below the larynx.*

An example is shown in Figure 6 using the MRI data from [10], which includes several cm of the trachea $r(x)$ below the glottis. Here, although both the magnitude and phase-based fits are reasonable in terms of length, since the acoustic length of that subject's subglottal tract has previously been reported as 195 or 216 mm using the frequencies of the measured impedance extrema and the calculations in [8] and [11]. However, the combined fit is not, showing that more work is needed.

4. Conclusions

For a rigid, dry vocal tract model closed at the glottis, seven or eight cylindrical elements seem to be the optimum model for $Z(f)$. Such a fit gives a rough representation of the shape on a resolution of about a centimetre, and the fitted $Z(f)$ through the lips matches very well. The fit for $Z(f)$ at the glottis is reasonable for the frequencies of its extrema, but not good for their magnitudes, due mainly to the strong influence of the epilaryngeal tube.

To fit *in vivo* measurements, it is necessary to increase the attenuation coefficient α to account for losses on wet, yielding tissues, as has been previously reported [7]. At low frequency, yielding walls introduce the tissue inertance and the tissue losses. These are included here as free parameters required to reproduce the extra low frequency minimum and maximum in $Z(f)$ measured through the lips.

In a preliminary trial, an open cylinder below the glottis was added to reproduce the extra extrema in Z observed when the glottis was open during inhalation, which allowed acoustic excitation of both vocal tract and trachea [8].

5. Acknowledgements

We thank the ARC for support and our volunteer subjects. This project was approved by the UNSW ethics committee.

6. References

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

[2] Mermelstein, P., "Determination of the Vocal-Tract Shape from Measured Formant Frequencies", J. Acoust. Soc. America 41, 1283 1967; doi: 10.1121/1.1910470Exam

[3] Ladefoged, P., Harshman, R., Goldstein, L., and Rice, L., J. "Generating vocal tract shapes from formant frequencies" Acoust. Soc. America 64, 1027, 1978, doi: 10.1121/1.382086

[4] Marshall, I. Rogers, M. and Drummond, G., "Acoustic reflectometry for airway measurement. Principles, limitations and previous work" Clin. Phys. Physiol. Meas. 12(2): 131-141, 1991.

[5] Story, B. H., Titze, I. R., and Hoffman, E. A., "Vocal tract area functions from magnetic resonance imaging" J. Acoust. Soc. America, 100(1),537-554, 1996.

[6] Almeida, A., Lehoux, H., Hanna, N., Smith, J. and Wolfe, J., "Estimating pressure and flow at remote locations in a vocal tract from microphone measurements elsewhere", Speech Science and Technology Conference, 2018

[7] Hanna, N., Smith J. and Wolfe, J., "Frequencies, bandwidths and magnitudes of vocal tract and surrounding tissue resonances, measured through the lips during phonation" J. Acoust. Soc. America, 139(5):2924–2936, 2016.

[8] Hanna, N., Smith J. and Wolfe, J., "Acoustic response of the subglottal tract measured by impedance spectrometry through the lips" J. Acoust. Soc. America, 143(5):2639-2650, 2018.

[9] Fletcher, N. H. and Rossing, T. D., The physics of musical instruments. Springer-Verlag, New York, 1998.

[10] Hanna, N., Amatoury, J., Smith, J., and Wolfe, J., "How long is a vocal tract? Comparison of acoustic impedance spectrometry with magnetic resonance imaging" Proc. Mtgs. Acoust. 28, 060001, 2016 doi: 10.1121/2.0000400

[11] Lulich, S.M., Alwan, A., Arsikere, H., Morton, J.R., Sommers, M.S., "Resonances and wave propagation velocity in the subglottal airways," J. Acoust. Soc. Am. 130(4):2108–2115, 2011.

Estimating pressure and flow at the glottis in a vocal tract-like duct from microphone measurements at the mouth

Hugo Lehoux, Andre Almeida, Noel Hanna, Joe Wolfe and John Smith

School of Physics, UNSW Sydney

a.almeida@unsw.edu.au

Abstract

We describe a method to deduce pressure and flow at the glottis from acoustic measurements near or downstream from the lips, using phonation into a cylindrical duct with three microphones. In this paper this method is tested on hard one-dimensional ducts, also reporting the precision of using 3D printed models of vocal tract area functions. Combining all the approximations yields results that are poor for the frequency range needed for phonetics, but they perform better at the frequencies of the first few voice harmonics.

Index Terms: vocal folds, physics, acoustic propagation

1. Introduction

One of the long-term goals of voice research is to calculate the pressure and flow at the glottis from measurements made through the lips without prior assumptions about the glottal flow. Several approaches to determining acoustic variables at the glottis of a speaker or singer have been reported: Rothenburg measured flow at the mouth using a mask [1] and Alipour [2] measured flow velocity and pressure directly on excised larynges of mammals using for instance a hot-wire probe, while others calculate these quantities using inverse filtering that requires some assumptions about the shape of the glottal flow [3,4], or a priori estimations of the vocal tract transfer function [5]. Inverse filtering has the advantage of being a non-invasive technique, but the assumptions about the glottal flow may be sometimes too strong for studies of the 3-dimensional motion of real vocal folds.

This paper suggests an alternative technique based on a method of acoustic impedance measurement that uses the pressure information provided by two or more microphones along a cylindrical tube (an impedance head) to deduce pressure and flow anywhere within that tube. The same formulae that are applied in that acoustic measurement can, in principle, be applied to determine the pressure and flow anywhere along the duct being measured, provided that the duct geometry is known and the frequencies of interest and the load geometry are such that the acoustic propagation can be considered as one-dimensional (1d). This in turn allows the determination of a (frequency-dependent) transfer matrix relating the Fourier magnitudes of pressures and flows at any two positions in the duct.

To test this approach, acoustic propagation in a known 1d duct is applied to the calculation of the flow and pressure in artificial ducts. The artificial ducts (*e.g.* Fig 1) have an area function $A(x)$ (cross section A as a function of distance x from the glottis) calculated from scans of real vocal tracts for given vowel articulations. 3D printed vocal tract models have been used to model the acoustic characteristics of vocal tracts using excitation at the glottis and to test measurements of transfer functions (e.g. [6,7,8,9]). Sometimes the quality of printing is an uncontrolled variable. In the experiments described here, known flow and pressure can be input at the 'glottis' end using one impedance head. This can be compared with the calculations using the propagation model and measurements made with a second impedance head at the 'lip' end. As such we can report on the acoustic quality of 3D-printed vocal tracts, benchmarked against some simple geometries. (We shall use 3D for the fabrication technique, and 3d for three-dimensional in other senses.)

For most of the measurements, a loudspeaker and a synthetic broad-band signal is used for spectral measurements. In other measurements, the voice itself or a synthesised glottal flow is used.

2. Materials and methods

2.1. 3D printed one dimensional tracts

1d vocal tract geometries were 3D-printed on a Cubicon Single Plus (3DP-310F) printer in PLA filament (Filaform and Verbatim 1.75 mm) based on previously published area functions calculated from MRI scans of real vocal tracts [10]. The area functions were adjusted slightly to match the lip end smoothly to the 26.2 mm diameter of one impedance head, and to the 7.8 mm radius at the glottis to match the other impedance head. Most tracts were printed in two longitudinal halves (see Fig. 1), in order to inspect the accuracy of the print. The two halves were later bonded together by applying an epoxy resin to the join between the two external surfaces to prevent liquid flowing into the airway.

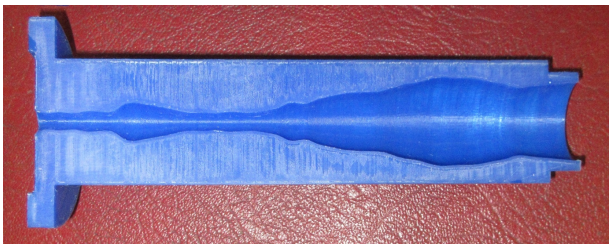


Figure 1: *Half of a 3D printed vocal tract model for the vowel /a/ from the data of Story et al. (1996). Note the left (glottis) side has been adjusted to fit the 7.8 mm diameter impedance head, and the right (lip) side has been adjusted to fit the 26.2 mm internal diameter impedance head.*

A few simple geometries such as cylinders and conical tract shapes were also printed and prepared in the same way and some were also printed in one piece for comparison.

2.2. Experimental setup

The experimental setup uses two cylindrical impedance heads as shown in Fig. 2. One head is used to generate a known acoustic wave at the 'glottis', and a second head at the 'mouth' measures the output wave and determines the pressure and flow at the glottis using only the output signal. The printed vocal tract is fitted between the two heads ($0 \ldots x_M$).

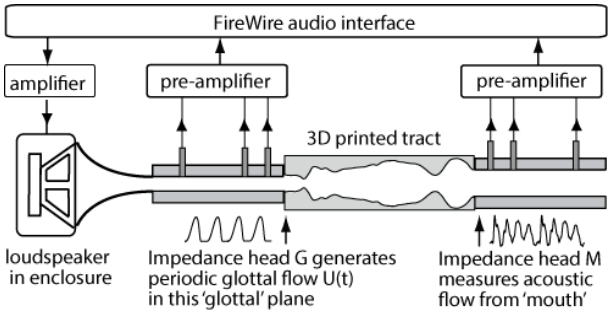


Figure 2. *Sketch of the experimental setup used for generation of known flow and estimation of wave propagation*

In a cylindrical duct, and for wavelengths long compared with the diameter, the formulae that relate pressure and flow at any two cross-sections of the duct are simple and accurate, and results can be improved by calibrating the impedance head using at least two different known acoustic loads (here a very high impedance provided by a 2.4 kg metal mass, and a very long pipe that has very delayed and very attenuated reflections.) The measuring 'mouth' head uses the 3-microphone technique described in [11]. After careful calibration, the measurements from the 3 microphones is used to determine the pressure and flow at x_M (see Fig. 2).

The generating 'glottis' head uses a principle similar to that described in [12]: After calibrating the microphones, the loudspeaker sends an initial glottal flow signal. The resulting signal is measured and after a series of iterations, it approaches the desired acoustic flow at the output $x = 0$. The target and measured flows agree to less than 0.1% as shown in Fig. 3.

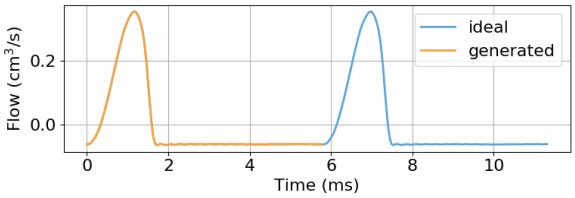


Figure 3: *Target flow pulse (blue) compared to obtained pulse (orange). After iteration, the maximum difference is less than 0.1%*

3. Results

3.1. Benchmark of printing

3.1.1. Accuracy of the print

Using digital photographs of the two halves and by visually recognising the inner edge of the tracts, the printed profiles were compared with the shapes that were provided to the printer and the original profiles. The two latter do not match exactly because the CAD software uses a spline interpolation to fill in the low-resolution original profiles. Fig. 4 shows an example of these measurements on one of the vowel models studied.

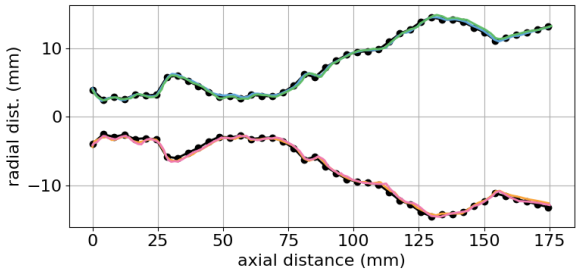


Figure 4: *Measurements of profile from printed two halves compared to the desired geometry*

The two profiles agree to 0.2 mm on average with a maximum difference of 0.5 mm (Fig. 4), an error which is estimated to be smaller than that due to the measurement of the profile in a digital photograph.

Although not clear from figure 1, the 3D printing process adds to the inner surface roughness of about 0.2 mm due to the thickness of the PLA filament. We chose not to polish this to avoid unintended changes to the inner geometry.

Another important factor is the finishing of the two ends of the tract that are used to connect to the impedance heads. The larger base consists of a flat surface that is somewhat polished and to which is applied petroleum jelly to provide a better seal. The opposite end connects to the impedance head via an adapter ring. Due to rugosity, some petroleum jelly is also needed to ensure an airtight seal around this ring.

3.1.2. Acoustic accuracy of printed tracts

The main aim of the 3D prints is to provide an accurate known acoustic system that can be used to test acoustic models. The most basic test is to make sure that a cylinder is accurately printed in order to mimic a cylindrical duct, here a PVC pipe 17cm long cut by hand (Fig. 5). The tract printed in a single piece shows slight changes in positions and magnitudes of resonances which are presumably due to roughness of the walls. After bonding, the 2-piece print has lower frequency resonances than the single piece print although there were no noticeable leaks at DC pressure.

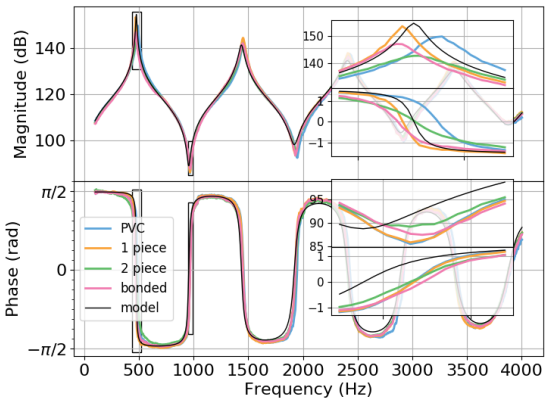


Figure 5: *Comparison of input impedance model and measurements for cylindrical straight ducts (a 17cm PVC pipe, 3D-printed in a single piece, printed in 2 halves before and after bonding). Insets are zooms.*

3.1.3. *Accuracy of the acoustic model of printed vocal tracts*

To test the modelling of changes in area function, a slightly more complex duct with a strong discontinuity was printed: it consists of a 170 × 7.8 mm diameter tube connected to a 26.2 mm tube, also 170 mm long. Figure 6 shows that the positions of secondary resonances (splitting of the full-length resonance due to discontinuity in cross-section) are not well captured by the model, indicating the need for a better model of acoustic impedance when discontinuities occur in the area function $A(x)$.

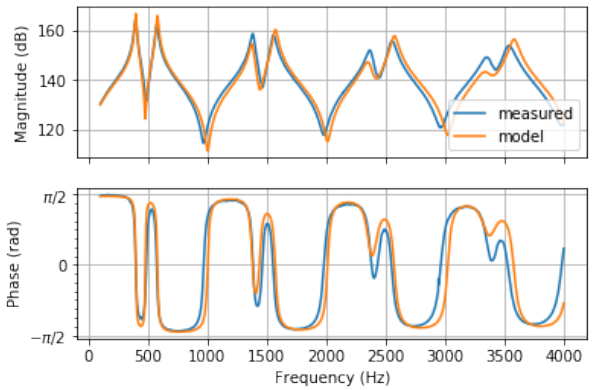


Figure 6: *Acoustic impedance of a cylinder 7.8 mm in diameter stepping into a cylinder 26.2 mm in diameter, compared with the model.*

1d computational models and 3D printed ducts using area functions from [10] usually fit well for lower frequencies (up to around 1kHz) with larger deviations in the 1-4 kHz range. One example, for vowel /a/, is shown measured from the glottis with an open mouth (Figure 7) and from the mouth with a closed glottis in Figure 8. The differences are usually seen in the Q factor of the resonances and seem more important when there are more sudden jumps in the area function.

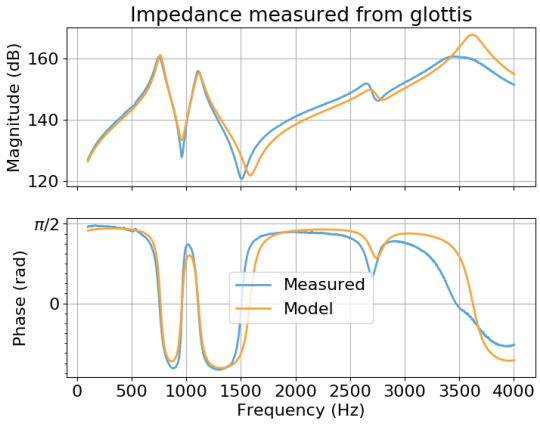


Figure 7: *The impedance measured on a 3D printed tract. Z at 'glottis' with an open 'mouth'; it is compared with a 1d model calculated with 100 segments*

3.2. Estimation of acoustic variables

In a later experiment, the signals recorded by microphones in the 'mouth' head were used to determine pressure and flow at x_M, and from these, using models of the propagation inside the tract with known geometry, the pressure and flow were estimated at $x = 0$.

For the vocal tract shown in Figures 1 and 4, the target flow and the estimated flow are shown in Figure 9, from which it is apparent that the lower frequency components are reasonably well estimated, whereas the higher frequency ones are much less accurate.

The agreement is fairly good for measurements at the fundamental frequency of the speaking voice, but by 700 Hz it is of the order of 3 dB with a phase difference of order $\pi/4$. Above 1 kHz the estimation is poorer with errors larger than 10 dB and phases that are almost uncorrelated with the target signal. These surprisingly large differences may be due to relatively rapid variations in $A(x)$ and thus the combination of multiple errors of the type seen in Figure 6. The poor performance at high frequency means that the technique is currently of little use for phonetics. However, for understanding the auto-oscillation of the folds and their interaction with acoustic loads at the fundamental frequency, it may have promise because it makes few strong assumptions.

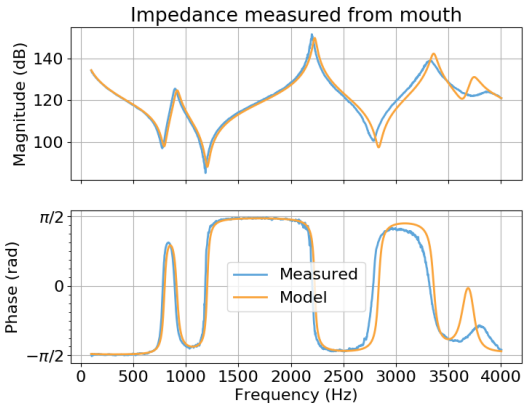


Figure 8: *As for Figure 7, but this time measured at the 'mouth' with a closed* 'glottis'

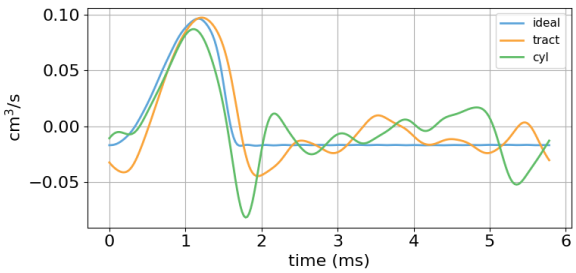


Figure 9: *Target flow pulse (blue) and estimated pulse from measurements at the mouth using propagation models based on the tract geometry (orange) or simply approximating it as a straight cylinder (green)*

3.3. Measuring downstream impedance from a glide

Given that the main purpose of this work is to deduce the acoustic wave generated by the vocal folds, the system was also tested in real-life conditions. It is the long-term aim to replace all the experimental apparatus upstream of the measuring impedance head with the mouth of a subject, where the $A(x)$ could be determined by the method described in the companion paper [13]. At the end of the mouth impedance head, a

loudspeaker is fitted that is used for a different experiment. In this preliminary experiment, it is shown that the human voice may be used to measure the impedance of the system downstream from the mouth.

The hardware setup of the impedance head is similar to the one in the previous experiment, however the analysis of the signals recorded by microphones 1-3M is slightly different, because the excitation signal is no longer strictly periodic.

If the singer performs a pitch glide that extends over at least one octave, there is, in principle, enough information in the microphone signals to extract the complex amplitudes at each frequency. From these, and given the same calibration data as for the previous measurement, it is possible to calculate pressure and flow of the plane wave travelling inside the impedance head. The ratio of pressure and flow gives the acoustic impedance of the passive system downstream of the head (Fig. 10). This information is important to determine the load acting on the vocal folds.

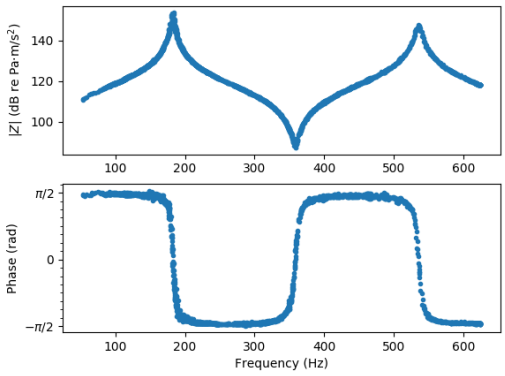


Figure 10: *Downstream impedance of one of the measurement heads open at the driver end, calculated from the first harmonic of a sung pitch glide*

4. Discussion

3D printing provides a quick and affordable way of producing test geometries that reproduce approximately the acoustic loads seen by the vocal folds at frequencies in the range of the fundamental f_0 of speech.

In order to produce 3D printed tracts that are useful for acoustic measurements, the ducts must be printed in a way allows quality control of the geometry. In particular, extraneous filaments should be removed, and the geometry of the print should be checked against the desired geometry. The disadvantage of this process is that the duct must be printed in two or more pieces and then glued together after polishing the contacting surfaces. Extra care needs to be given to ensure the airtightness of the two halves after gluing and the fit of the duct to the measurement device, for example by using grease to seal the contact. Air leaks decrease the magnitude of measured impedance peaks.

In these conditions, the acoustic properties below 1 kHz are relatively well predicted by a simple propagation theory that considers 1 dimensional plane waves (*e.g.* Figs 6-8). Geometries with rapidly varying cross-sections are particularly challenging and will probably need a more sophisticated propagation model. Using the current propagation model, calculations that treat the (complicated) duct as a simple cylinder provide results whose accuracy compares with that calculated using the real geometry. However, simpler experimental geometries such as straight ducts and slowly varying cross-sections provide an extended range of usable frequencies up to 4 kHz, when calculated with the real geometry. These results point towards improving the propagation model at discontinuities using lumped elements that include an inertive end effect and a stagnant air compliance.

The calculations of the transfer matrices can be compared with the use of inverse filtering, e.g. in [4]. In that work, the errors in the estimation of the glottal flow are significantly smaller than in the current study, although in that study the injected flow follows very closely the assumptions of the inversion method, in particular the frequency content of the flow signal follows the model used in the inversion method.

5. Acknowledgements

We thank the Australian Research Grants Council for support.

6. References

[1] Rothenberg, Martin. “A new inverse-filtering technique for deriving the glottal air flow waveform during voicing.” The Journal of the Acoustical Society of America 53(6): 1632-1645, 1973.

[2] Alipour, F., and Scherer, R. C., “Pulsatile airflow during phonation: an excised larynx model” J. Acoust. Soc. America, 97(2): 1241-1248, 1995.

[3] Airas, M., “TKK Aparat: An environment for voice inverse filtering and Parameterization,” Logopedics Phonatrics 33(1): 49–64, 2008.

[4] Chu, D.T.W., Li, K., Epps, J., Smith, J. and Wolfe, J., “Experimental evaluation of inverse filtering using physical systems with known glottal flow and tract characteristics” J. Acoust. Soc. America. 133(5): EL358-362, 2013.

[5] Alku, P. “Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering.” Speech communication, 11(2-3), 109-118, 1992

[6] Kitamura T, Takemoto H, Adachi S, Honda K. “Transfer functions of solid vocal-tract models constructed from ATR MRI database of Japanese vowel production.” Acoust. Sci. Tech. Tech. 30(4):288–296, 2009.

[7] Delvaux B., and Howard D. “A New Method to Explore the Spectral Impact of the Piriform Fossae on the Singing Voice: Benchmarking Using MRI-Based 3D-Printed Vocal Tracts,” PLOS ONE 9(7): e102680, 2014. https://doi.org/10.1371/journal.pone.0102680

[8] Echternach M., Birkholz P., Traser L., Flügge T. V., Kamberger R., Burk F., et al. “Articulation and vocal tract acoustics at soprano subject’s high fundamental frequencies.” J. Acoust. Soc. America 137(5):2586–2595, 2015. pmid:25994691.

[9] Fleischer, M., Mainka, A., Kürbis, S. and Birkholz, P. “How to precisely measure the volume velocity transfer function of physical vocal tract models by external excitation” PLOS ONE 13(3): e0193708, 2018.

[10] Story, B. H., Titze, I. R., & Hoffman, E. A. “Vocal tract area functions from magnetic resonance imaging.” J. Acoust. Soc. America, 100(1): 537-554, 1996.

[11] Dickens, P., Smith, J. and Wolfe, J. (2007) “Improved precision measurements of acoustic impedance spectra using resonance-free calibration loads and controlled error distribution”, J. Acoust. Soc. America, 121(3): 1471-1481.

[12] Wolfe, J., Chu, D., Chen, J.-M. and Smith J., “An experimentally measured Source-Filter model: glottal flow, vocal tract gain and radiated sound from a physical model”, Acoust. Australia, 44: 187–191, 2016.

[13] Hanna, N., Rodriguez, A., Almeida, A., Wolfe, J., Smith, J., “Estimation of vocal tract and trachea area functions from impedance spectra measured through the lips”, Speech Science and Technology Conference, 2018

Acoustic characteristics of pre-aspiration in Australian English

Simon Gonzalez

ARC Centre of Excellence for the Dynamics of Language, Australian National University
simon.gonzalez@anu.edu.au

Abstract

This study analyzed the acoustic characteristic of pre-aspiration in Australian English. We specifically examined the laryngeal coordination from vowel to coronal obstruents in 15 Australian English speakers from the Sydney area. The analysis was based on four acoustic parameters: duration, Centre of Gravity, Spectral Tilt, and previous vowel formant frequencies. Results showed that pre-aspiration is attested in the data in 68% of the cases. This serves as a starting point for more robust analysis of pre-aspiration in this English variety as well as its potential sociolinguistic implications.

Index Terms: Australian English, coronal obstruents, pre-aspiration

1. Introduction

Pre-aspiration is defined as a period of voiceless glottal friction which can occur before voiceless obstruents, both stops and fricatives, and is usually preceded by a breathy transition from the vowel [1,2]. This has been observed and documented in different English varieties, including Middlesbrough English [3], Newcastle English [4], Scottish Gaelic [5], and Welsh English [2]. Pre-aspiration been observed only rarely in Australian English [1,2]. Studies on Australian English have mainly reported frication and lenition of /t/ [1,6]. In [1], pre-glottalization was regularly attested by one Australian English speaker. This glottalization is classified as one type of pre-aspiration in [2]. Sporadic findings of pre-aspiration are also observed in [7,8,9], in which single speakers show this feature. One study looking at patterns in the acoustics of /p, t, k/ in children native speakers of Australian English found that pre-aspiration is attested by female speakers, but the rate of use decreases as age increases [10]. These findings therefore suggest that even when it is not consistent or systematic as in studies in other varieties, pre-aspiration needs closer inspection in Australian English.

Acoustic studies on pre-aspiration have shown that word position and previous vowel significantly impact duration and frequency of pre-aspiration, as in the case of Scottish Gaelic in which pre-aspiration is a well-established feature in the phonological structure of the language [11]. In terms of word position, word-final stops prompt longer pre-aspiration durations than word-medial stops. In terms of the preceding vowel, only vowel height showed significant differences in the pre-aspiration duration, in which non-high vowels trigger longer pre-aspiration durations than high-vowels.

Another relevant finding in [11] was the existence of a breathy voice component at the end of the preceding vowel in one-third of the tokens with pre-aspiration. This shows that pre-aspiration also affects the realization of the previous vowel by making the vowel breathy at the end of its realization, which is the result of the same glottal abduction gesture. A distinction therefore can be made within the vowel between modal voicing and non-modal voicing. In [3], important coarticulatory correlations were observed between the pre-aspirated section and its impact on the previous vowel. They report vocal fold abduction extended in time, which is suggestive that pre-aspiration should be analyzed as an articulatory property of the voiceless coda segment and not exclusively as a part of the plosive articulation itself. This is in line with what is reported in [2] where it was evidenced that breathiness is an articulatory transitional section of the vowel.

Place of articulation has also been reported to have a significant impact on pre-aspiration. In [5], place of articulation of the target segment showed significant effects on pre-aspiration duration, with labials triggering shorter durations than coronals and dorsals. This is in contrast with [11] in which no significant differences were found between place or articulation. However, results in [11] showed that coronal segments are more likely to trigger pre-aspiration than labials and dorsals.

Another important correlation found in previous studies is with gender. In [11], female speakers pre-aspirated with more frequency than males and in [5] it is stated that pre-aspiration tokens are allophonic realizations identified mainly in female speakers.

In this study, we examine pre-aspiration in Australian English. We aim to provide quantitative evidence of pre-aspirated coronal obstruents.

2. Methodology

We limit this analysis to word-final contexts, since they correlate with longer pre-aspiration sections, as in Gaelic [11]. We also present evidence on segments preceded by the open vowel /æ/ in Australian English, which also correlates with longer pre-aspiration section compared to high vowels [11].

2.1. Participants

A total of 15 native speakers of Australian English (all females) took part in this study. They were all first-year linguistics students at Macquarie University. None of the participants reported hearing or articulatory impairment.

2.2. Stimuli

We analyzed pre-aspiration in the English voiceless coronal obstruents /tʃ, t, s, ʃ, θ/ in coda position. Segments were placed in the carrier words, *catch*, *cat*, *Cass*, *cash,* and *Kath*, respectively. These words were chosen to have the same phonological context for all words, i.e /kæC/, in which C represents the target segment. All target segments were therefore preceded by the same vowel /æ/.

2.3. Procedure

Speakers were recorded in a sound-attenuated testing room and exposed to the target words in elicitation sentences, pre-recorded by a native speaker of Australian English. Sentences were presented to the participants in a Microsoft PowerPoint slide show in a monitor in front of them. Speakers were then asked to listen then repeat each sentence five times. Each target word was placed in the carrier sentence *Please, utter X*, where *X* represents the target word. A Shure KSM137 unidirectional microphone was used to record the acoustic signal. The audio signal was digitized at a sampling frequency of 48 KHz with 16-bit quantization.
Recordings were then forced-aligned using the Montreal forced-aligner [12] and then manually corrected by the author. The acoustic information was extracted in Praat [13] and the analyzed in Rstudio [14]. Out of the total 379 tokens available, 19 tokens were excluded from the analysis for errors in the original recordings, resulting in a total of 360 tokens. Token numbers are summarized in Table 1.

Table 1. Summary of number of tokens

Segment	Total tokens	Pre-aspirated tokens	Pre-aspirated percentage
tʃ	62	25	40.3
t	65	38	58.5
s	78	60	76.9
ʃ	78	61	78.2
θ	77	67	87
Total	360	251	68.19 (Mean)

2.4. Segmentation

Challenging issues permeate the literature on pre-aspirated consonants. Here we follow the parameters as in [2,11] in which the acoustic dynamic range was reduced to 50-65dB. This helped the distinction between pre-aspirated sections and background noise. Since pre-aspiration varies depending on the phonological environment (stops vs fricatives), we followed different protocols for each.

2.4.1. Vowel segmentation

The onset of the vowels was placed in the beginning of visible formants in the spectrograms and changes in the waveform coming from the frication of onset /k/. Vowel offset before non-pre-aspirated stops was defined by the cessation of vocalic pulses. The offset of the vowel before pre-aspirated segments was placed at the beginning of aperiodic high frequency energy belonging to the pre-aspirated section or the segment in the case of fricatives. Within the vowel, we also coded for breathiness (see *BR* section in Figure 1). The onset of breathiness was located at the end of modal voicing, which gives way to the breathiness section. The breathy section is distinguished from the pre-aspirated section by the presence of formant information from the vowel as well as partial vowel-like cycles in the waveform, as done in [2].

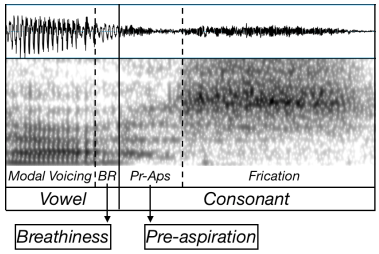


Figure 1. Acoustic cues for the segmentation of the previous vowel and the fricative consonant /s/

2.4.2. Fricatives segmentation

For the onset of fricatives /s, ʃ, θ/, the acoustic cue was spectral changes, where a distinction between high and low energy was observed (See Figure 1). The pre-aspirated section was generally distinguished from the main frication section (see *Frication* section in Figure 1) in which the frication section had major concentrations in high-frequency energy.

2.4.3. Stops segmentation

In this study, we refer to both /t, tʃ/ as stop segments. Onset of pre-aspiration in stops was identified as the presence of high energy after the vowel and before complete closure (See *Pr-Asp* in Figure 2). The offset of the pre-aspiration section was identified at the point where the high-frequency energy disappeared to give way to the closure section of the stop segment. We also calculated the segment duration, from onset of *Closure* to the offset of the *Release* section.

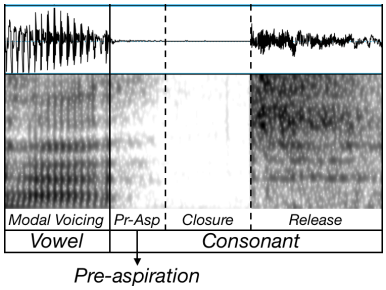


Figure 2. Acoustic cues for the segmentation of the previous vowel and the stop segment /t/

2.5. Measurements

We carry out our measurements following methodologies on previous studies. Four main parameters have been examined in the literature: pre-aspiration duration duration, Center of Gravity (CoG), vowel spectral tilt and formant frequencies of the previous vowel. In our measurements, we treat the breathiness section of the vowel as separate from the pre-aspirated section as suggested in [2]. This is different from other studies in which these two sections are treated as a single unit [15]. The former approach (separate sections) is based on evidence supporting pre-aspiration being phonetically conditioned by vowel height and length, whereas breathiness is sensitive to phonetic vowel duration [2].

2.5.1. Duration

In [3,16] it was shown that there is variation in the total duration of the previous vowel depending on the place and manner of articulation of the following pre-aspirated segment.

The total duration of the vowel was calculated as well as the breathy section to examine their relation to the pre-aspiration section. We analyzed whether the pre-aspirated consonants had an effect on the duration of the previous vowel.
We also calculated whether the durations of the pre-aspirated sections (*Pr-Asp* in Figures 1 and 2) were significant different based on *Place* and *Manner of Articulation*. Finally, we tested the relationship between the pre-aspirated sections and the duration of the segment without the pre-aspirated section, that is, the *Frication* section for the fricatives and the *Closure+Release* sections for the stops.

2.5.2. Centre of Gravity

Centre of gravity (CoG) is an acoustic measure that allows parameterization of the pre-aspirated section. CoG can measure the change within fricative productions [17]. This primarily correlates with the front cavity size and constriction shape; i.e. smaller front cavities have higher COG and larger front cavities have smaller CoG. This allows us to examine whether the front cavity size has an impact on the presence/absence or length of the pre-aspirated section. A Praat script was written to calculate the CoG for all the pre-aspirated sections, both in fricatives and stops.

2.5.3. Spectral Tilt

Spectral tilt (H1-H2) is an acoustic parameter that measures breathiness or creakiness in vowels [16]. This was calculated for each vowel to examine the effect of the pre-aspirated section in the preceding vowels. Spectral tilt was calculated using a Praat script by [18]. Higher H1-H2 values correlate with breathier vowels.

2.5.4. Formants of the previous vowel

Previous studies have shown that pre-aspiration has an impact on the previous vowel F1 values [16], in which F1 values were significantly different between aspirated and non-aspirated consonants (the direction of the significance is not reported, i.e. whether pre-aspirated F1 were higher or lower). We extend this measurement to include F2 values in addition to F1 values. These were calculated at the midpoint of the preceding vowel.

2.6. Statistical analysis

For all comparisons, we applied mixed effects models, with the analyzed values as the response variable. We created our models to check differences based on manner, place of articulation and duration (fixed factors), and speaker as a random factor. For manner of articulation we distinguished stops vs fricatives, and for place of articulation between palato-alveolars, alveolars, and dental.

3. Results

3.1.1. Duration

Table 2 summarizes the mean durational values (in ms) and corresponding total tokens for phonological context in all pre-aspirated tokens.

Table 2. Summary of means and standard deviations (in parenthesis) of durational values

Context	Segment *ms*	Pre-aspiration *ms*	Previous vowel *ms*
Fricative [188]	291 (56)	60 (19)	143 (26)
Stop [63]	259 (61)	52 (21)	123 (22)
Alveolar [98]	275 (62)	60 (20)	139 (26)
Dental [67]	265 (45)	57 (18)	140 (27)
Pal.-alv. [86]	307 (56)	57 (22)	136 (28)

Vowel duration analysis showed that there was a significant effect for manner ($\chi2(1)=38.227$, $p<0.001$, shorter durations by 20.3 ms ± 3.2 for vowels followed by stops), and a marginally significant difference for place of articulation ($\chi2(2)=5.029$, $p=0.081$). With alveolars as the baseline, the dental had shorter durations by 7.4 ms ± 3.3, and palato-alveolars by 3 ms ± 2.9.
Duration analysis comparing the pre-aspirated sections also showed significant differences. Manner of articulation was found to be a predictor between segments ($\chi2(1)=6.088$, $p=0.014$, shorter durations by 7.009 ms ± 2.9 for stops). There were no significant differences for place of articulation.
The analysis comparing the pre-aspirated section of the consonant with the duration of the segment also revealed significant differences. Overall, the pre-aspirated section accounted for 27.19% of the total duration of the segment. The duration of the pre-aspirated sections were significantly different than the other section of the segments ($\chi2(1)=1028$, $p<0.001$, shorter durations by 166.79 ms ± 2.8 for the pre-aspirated section).

3.1.2. Centre of Gravity

CoG analysis showed significant differences between the pre-aspirated sections across all segments. Analysis showed that place of articulation was the only factor to be a stronger predictor. The only significant difference was between the alveolar and the palato-alveolars ($t=2.390$, $p=0.0176$, higher CoGs by 55.4 Hz ± 23.18 for the palato-alveolars). There were no significant differences for the other comparisons.

3.1.3. Spectral Tilt

Spectral tilt analysis did not show significant differences. However, average values showed that vowels followed by the dental segment had breathy sections with the highest energy (mean = 4.59 ± 5.8 dB), vowels followed by alveolars having the less breathy segments (mean = 2.86 ± 6.6 dB), and the palato-alveolar context in between (mean = 3.49 ± 6.6 dB).

3.1.4. Formants of the previous vowel

There were no significant differences in the formant analysis. However, F1 analysis showed higher values for vowels followed by a pre-aspirated segment (mean = 984 ± 142 Hz) than non-aspirated tokens (mean = 960 ± 114 Hz). F2 analysis showed lower values for vowels followed by a pre-aspirated segment (mean = 1959 ± 237 Hz) than non-aspirated tokens (mean = 1982 ± 223 Hz).

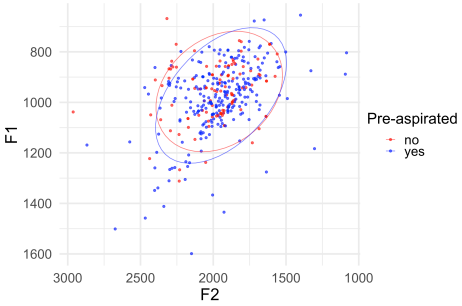


Figure 3. F1/F2 vowel space for the previous vowel in relation to following segment, whether it was pre-aspirated or not

4. Discussion

This study tested the presence of pre-aspiration in Australian English. We examined laryngeal coordination in voiceless coronal obstruents based on four acoustic parameters: duration, CoG, spectral tilt and formants of the previous vowel.

Durational analyses reveal that both manner and place of articulation of the pre-aspirated segment affect the duration of the previous vowel, with alveolar fricatives triggering longer durations. Results also show that fricative segments favor longer vowel durations than stop segments. Finding significant results in place of articulation is in line as in [5], and in contrast to [11] in which there were no significant differences in place of articulation.

Centre of gravity on the pre-aspirated section analysis showed that it can be a reliable acoustic cue for place of articulation, with alveolars having the lowest CoG values and palate-alveolar the highest. This is in contrast to CoG analysis on the frication section of segments, in which palato-alveolar have lower values than alveolars, due to alveolars having shorter front cavities. This suggests that, as the vocal tract transitions from the vowel to the pre-aspirated section of the consonant, this transition has its own parameters of articulation. This requires further analysis. Spectral tilt results suggest that the dental segment produces more breathiness in the previous vowel than palato-alveolars and alveolars, with alveolars producing the least amount of breathiness in the previous vowel.

Considering these three analyses together, results show that pre-aspiration in alveolars affects mainly the duration of the previous vowel, whereas the dental and the palato-alveolars affect mainly the breathiness of the previous vowel. Finally, formant analysis does not show to be a reliable parameter related to pre-aspiration. However, results show that pre-aspiration produces slightly more open and more retracted vowel midpoints. Further analysis can be carried out to test this finding with a wider range of vowels.

5. Conclusions

Evidence presented in this study confirms the existence of pre-aspiration in coronal obstruents in Australian English, in female speakers from the Sydney area. Though not systematic, pre-aspiration is observed in most of the cases examined. However, further work is needed to test whether any of the tokens are heard as pre-aspirated. This study therefore provides new evidence on a phonetic feature previously analysed in other English varieties but not systematically analysed in Australian English. More research is needed to further examine pre-aspiration in more contexts and its potential sociolinguistic implications in Australia.

6. References

[1] Jones, M.J. and McDougall, K. "The acoustic character of fricated /t/ in Australian English: a comparison with /s/ and /f/", Journal of the International Phonetic Association 39(3): 265289, 2009

[2] Hejná, M. "Pre-aspiration in Welsh English: A case study of Aberystwyth", PhD Dissertation, University of Manchester, 2015

[3] Jones, M. and Llamas, C. "Fricated pre-aspirated /t/ in Middlesbrough English: An acoustic study", in 15th International Congress of Phonetic Sciences, Barcelona, Universitat Autónoma de Barcelona (ICPhS15), 655–658, 2003

[4] Docherty, G. and Foulkes, P. "Instrumental phonetics and phonological variation: Case studies from Newcastle upon Tyne and Derby", In Paul Foulkes & Gerard Docherty [Eds], Urban voices: Accent studies in the British Isles, 47–72. London: Arnold, 1999

[5] Nance, C. and Stuart-Smith, J. "Pre-aspiration and post-aspiration in Scotish Gaelic stop consonants", Journal of the International Phonetic Association, 43, pp 129-152, 2013

[6] Loakes, D. and McDougall, K. "Individual Variation in the Frication of Voiceless Plosives in Australian English: A Study of Twins' Speech", Australian Journal of Linguistics, 30:2, 155-181, 2010

[7] Horvath, B.M. "Variation in Australian English: the sociolects of Sydney Cambridge", Cambridge University Press, 1985

[8] Ingram, J.C.L. "Connected speech processes in Australian English", Australian Journal of Linguistics 9: 2149, 1989

[9] Jones, M.J. and McDougall, K. "A comparative acoustic study of Australian English fricated /t/: assessing the Irish (English) link", in P Warren & CI Watson [Eds] Proceedings of the 11th Australasian International Conference on Speech Science and Technology 68 December 2006 Auckland: Australasian Speech Science and Technology Association. pp. 612, 2006

[10] Tait, C. and Tabain, M. (2016). "Patterns of gender variation in the speech of primary school-aged children in the Australian English: the case of /p t k/", in 16th Australasian International Conference on Speech Science and Technology. Sydney: Australia, 65-68.

[11] Clayton, I. "Preaspirated stops in the English of Scottish Gaelic-English bilinguals", in The Scottish Consortium for ICPhS 2015 (ed.), Proceedings of the 18th International Congress of the Phonetic Sciences, August 10-14, 2015. Glasgow, Scotland: the University of Glasgow, 2015

[12] McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M. and Sonderegger, M. "Montreal Forced Aligner: trainable text-speech alignment using Kaldi", in Proceedings of the 18th Conference of the International Speech Communication Association, 2017

[13] Boersma, P. and Weenink, D. "Praat: doing phonetics by computer [Computer program]", version 6.0.40, retrieved 11 May 2018 from http://www.praat.org/, 2018

[14] RStudio Team. "RStudio: Integrated Development for R", RStudio, Inc., Boston, MA URL http://www.rstudio.com/, 2015

[15] Dommelen, W. A. van, Holm, S. and Koreman, J. "Dialectal feature imitation in Norwegian". 17th ICPhS. Hong Kong, 599–602, 2011

[16] Sarmah. P. and Mazumdar, P. "Aspiration in alveolar fricatives in Bodo", ICPhS 2015At: Glasgow, 2015

[17] Nittrouer, S., Studdert-Kennedy, M. and McGowan, R.S. "The emergence of phonetic segments: Evidence from the spectral structure of fricative-vowel syllables spoken by children and adults", Journal Speech and Hearing Research, vol. 32, pp. 120-132, 1989

[18] DiCanio, C. "Spectral Tilt Script for Praat", Praat Script retrieved from http://www.acsu.buffalo.edu/~cdicanio/scripts.html, 2007

Rosa's roses – unstressed vowel merger in Australian English

Felicity Cox and Sallyanne Palethorpe

Centre for Language Sciences, Department of Linguistics, Macquarie University

felicity.cox@mq.edu.au; sallyanne.palethorpe@mq.edu.au

Abstract

Contrast loss between unstressed vowels in words like *Rosa's*/*roses* is assumed for Australian English but no empirical analyses have been conducted to support this assumption. To address this lacuna we acoustically examined the unstressed vowels in word-final (e.g. *Rosa*), segmental possessive (e.g. *Rosa's*), and syllabic plural (e.g. *roses*) contexts for 27 female Australian English speakers. Results showed that word-final schwa differed in F1, F2 and duration from the other types. Unstressed vowels in segmental possessive and syllabic plural contexts did not differ significantly from one another. Results support the merger of weak vowels /ə/ and /ɪ/ for Australian English.

Index Terms: Australian English, schwa, weak vowel merger, unstressed vowels

1. Introduction

According to [1], Australian English (AusE) has a relatively high occurrence of schwa (28.4% of all vowels) compared to Received Pronunciation (RP) (23.7%). In contrast, /ɪ/ makes up just 11% of vowels in AusE compared to 20.8% in RP. One of the reasons for this discrepancy is that /ə/ and /ɪ/ are said to have undergone a merger in unstressed syllables in AusE whereas in RP and Southern Standard British English (SSBE) the two vowels remain distinct. [2] refers to this loss of contrast between /ə/ and /ɪ/ as Weak Vowel Merger. Thus, contrast between the weak vowels in words such as *Lennon* vs *Lenin* has reportedly been lost in AusE. /ɪ/ is typically retained in suffixes like *-ish, -ic, ism, ing*, and /ɪ/ often still occurs when the following consonant is postalveolar or velar but this could possibly be considered allophonic (or free variation) e.g. *paddock, stomach, beverage, manage.* To explain the development of Weak Vowel Merger, [3] suggests that the dialects present in the early European colony in Australia would have included speakers from Ireland, East Anglia, The West Country and the far north of England for whom schwa would occur in unstressed syllables in words like *rabbit*. Speakers from Scotland, the London region and central areas of England would have used /ɪ/ instead. [3] suggests that speakers of the merged forms would have been in the minority in the early Australian European colony, yet a merger was ultimately adopted by the children through a process of simplification that occurs in dialect contact situations leading to the survival of those minority variants displaying greater regularity than the majority variants ([4], [5], but see [6] for an alternative explanation). New Zealand English also displays Weak Vowel Merger leading [7] to conduct an auditory analysis of the unstressed vowel in the *rabbit* class of words produced by 59 speakers from the historical ONZE corpus who were born in the latter half of the 19th century. After coding tokens on a three-point scale of centralisation, [7] found that 16.6% of speakers used the schwa variant, 32.4% used a retracted variant of [ɪ], with the remaining 51% using fronted [ɪ]. They also found an increase in the use of schwa over time, particularly for female speakers, which they suggest demonstrates change in progress. Speakers with the greatest use of schwa were of Irish descent indicating a potential Irish English influence and the possibility that Weak Vowel Merger may be a phonological remnant of Irish English in Southern Hemisphere varieties today.

In one of the few acoustic studies of unstressed vowels in AusE, [1] examined schwa production in the speech of 20 males from Rockhampton and Sydney. Speakers read 18 sentences containing schwa in a range of contexts including utterance initial (indefinite article), utterance final (non-rhotic *-er* suffix) and medial interconsonantal (indefinite article preceded by a word containing a coda stop and followed by a word containing an onset stop - C_1#ə#C_2). The results showed that utterance final /ə/ was positioned lower in the phonetic space (raised F1) than medial schwa, and that horizontal displacement (F2 variation) was determined by surrounding consonantal context. [8] analysed the unstressed vowel in trochaic words of /(C)VɹəC/ type such as in *parrot, syrup, barrack* in the speech of 28 females from Sydney and found F2 to be significantly affected by place of articulation of the following stop but that F1 was not affected by this feature. [9] suggests that extreme contextual variability in schwa is due to its short duration, explaining variability with reference to the undershoot model.

Schwa has been described as 'targetless' and entirely assimilated to its phonetic context ([11]). Coarticulation affects unstressed vowels more than accented vowels ([12]) so it is not surprising that schwa is highly susceptible to its environment. [13], based on an articulatory study of a single speaker of American English, described schwa as an average of the articulatory positions of the full vowels (articulated towards the centre of the vowel space) and a goal driven gesture therefore one with a specified target rather than being targetless. In response, [14] tested the hypothesis that schwas contained within the syllabic past tense affix might arise through coordination of the consonants on either side rather than any kind of targetful phonologically specified vowel. They used real-time MRI to detail the posture, duration, acoustic and articulatory targets of lexical schwa and the schwa in past tense affixes (e.g. *Nota'd* vs *noted* – where *Nota'd* was a contraction of 'Nota had') in the speech of two male speakers of American English. Their findings showed significant differences in both acoustic (F1/F2) and articulatory parameters between lexical schwa and affix schwa but also indicated that the schwa could not be considered targetless. [15] also suggested targetful schwa after conducting an analysis of X-ray data from four speakers of American English showing significant retraction of the tongue root during schwa relative to resting posture.

In a similar finding to that of [1], [16] found wide variation in F1 of word-final schwa which was phonetically lower than non word-final schwa in the speech of 9 female speakers of American English. They suggest that phonetically low word-final schwa is necessary to preserve the contrast with other

vowels such as /iː/ and /əʉ/ in unstressed contexts (e.g. in words like *pita, pity, ditto*). However, in non word-final contexts (e.g. in syllabic plural /-əz/) there are no contrastive competitors and hence no need for lowering as a differentiating strategy.

[10] examined F1, F2 and duration of word-final schwa extracted from conversational speech of 11 male and 10 female AusE speakers from Sydney divided into four ethnic groups (Greek, Lebanese, Italian, Anglo). A maximum of 30 words per speaker were selected. His results showed considerable variability in keeping with the uncontrolled phonetic nature of the extracted data, but a relationship was found between length and phonetic openness with longer schwa being more open.

Interestingly, [16] showed a significant height difference between the unstressed vowel in the syllabic plural (affix) (e.g. *roses*: lower F1 – phonetically higher) compared to the segmental plural and possessive forms of schwa-final words (*sofas/Rosa's*: higher F1 – phonetically lower). [16] assert that word-final unstressed vowel qualities are preserved under affixation of the segmental plural or possessive morpheme (e.g. *sofas/Rosa's*), hence their lowered realisation compared to the syllabic plural (affix) type (e.g. *roses*). [14] found a similar height difference in lexical vs affix schwa in their pair of American English speakers.

The contrast between raised and lowered unstressed vowels in pairs like *Rosa's* vs. *roses* is said to arise because of the difference in morphological structure and position of the stem boundary between the words: in *Rosa's* it follows the vowel (/ɹəʉzə|z/), while in *roses* the stem boundary precedes the unstressed vowel (/ɹəʉz|əz/). This morphological effect has not previously been explored for AusE. However, as AusE is said to have undergone Weak Vowel Merger, if a true merger exists we expect that the morphological effect found in [14] and in [16] will not hold. With this proposition in mind, we examined schwa in pre-pausal word-final position (e.g. *Rosa*), in morpheme final position before the segmental possessive (e.g. *Rosa's*), and in the syllabic plural (affix) /-əz/ (e.g. *roses*).

2. Method

Twenty-seven female speakers of AusE (18-38 years) with a mean age of 24 who had all been born in Australia to parents whose first language was English read 133 isolated words containing schwa in a range of contexts presented orthographically and individually on a computer screen three times in random order. The task also included an additional set of AusE vowels in the /hVd/ frame read three times in random order. Recordings were made in a sound treated recording studio in the Macquarie University Linguistics Department, using an AKG C535 EB microphone, Cooledit audio capture software via an M-Audio delta66 soundcard to a Pentium 4 PC at 44.1kHz sampling rate. We extracted sound files from each speaker for three repetitions of 12 target trochaic words (see Table 1) in addition to their /hVd/ words. 45 of the possible 972 tokens were excluded through errorful production. Data were first processed by the WebMAUS automatic aligner [17] utilising an AusE model. MAUS automatically returns Praat [18] textgrids with phonemic boundaries labelled. These were then checked and hand corrected where necessary to delimit the segmental boundaries in Praat [18] and Emu (ips-lmu.github.io/EMU.html). F1 and F2 were extracted for the /hVd/ items and for schwa at the midpoint. Duration of schwa was also determined. Note that the /hVd/ monophthongs are used for illustrative purposes only here to situate the schwa in the vowel space. F1, F2 and duration of the unstressed vowels were examined separately. Linear mixed model analysis was conducted in SPSS with fixed factors word group (ROSA, NINJA, LISA, ASIA), schwa type (word-final, segmental possessive, syllabic plural (affix)), repetition, and the interaction of word group and schwa type. As repetition did not show any significant interactions with the other two factors, this was only included as a main effect in the model. Speaker was included as a random factor. Post-hoc analyses with Bonferroni correction were carried out.

Table 1. *Single words extracted from acoustic recordings*

Word group	Schwa type			
	Word-final	Segmental possessive	Syllabic plural-affix	
ROSA	*Rosa*	*Rosa's*	*roses*	/-zəz/
NINJA	*ninja*	*Ninja's*	*hinges*	/-ʤəz/
LISA	*Lisa*	*Lisa's*	*leases*	/-səz/
ASIA	*Asia*	*Asia's*	*ages*	/-ʒəz/-/-ʤəz/

3. Results

The results show significant effects for word group and schwa type and significant interactions for each of F1, F2 and duration (Table 2). For F1 and F2, post-hoc analyses showed that word-final schwa differed from the other two types (p<0.001) for all word groups and that the possessive and syllabic plural (affix) did not differ from each other for any word group except ASIA (F1: p<.013, F2: p<.001) (see Figure 1). This F1 and F2 difference for ASIA can be explained by the difference in the preceding consonant between the two types (/ʒ/ - *Asia's* vs /ʤ/-*ages*). Figure 1 shows the F1/F2 ellipses (enclosing two standard deviations from the mean) for the four word groups for each of the unstressed vowel types and shows the ASIA (*Asia's* vs *ages*) effect in the bottom right panel. It also shows that word-final schwa is clearly separated spectrally from the interconsonantal schwa contexts. The possessive and syllabic plural schwa did not differ from each other in duration but both differed significantly from word-final schwa (p<.001) (see Figure 4). Word-final schwa was longer, lower and more retracted than the other contexts.

Post-hoc analyses showed that word groups differed from each other for each parameter in keeping with coarticulatory effects associated with differing segmental context. Figures 2, 3 and 4 illustrate the F1, F2 and durations for schwa type within each word group.

Table 2. *Summary of significant main effects and interactions*

Parameter	Variable	*df*	*F*	*p*<
F1	Word group	3,888	3.874	.009
	Schwa Type	2,888	3765.256	.001
	Word Group x Schwa Type	6,888	8.939	.001
F2	Word group	3,888	153.190	.001
	Schwa Type	2,888	1472.224	.001
	Word Group x Schwa Type	6,888	5.698	.001
Duration	Word group	3,888	22.697	.001
	Schwa Type	2,888	165.741	.001
	Word Group x Schwa Type	6,888	2.919	.008

For word-final context, LISA was significantly phonetically lower than the other three word groups (higher F1) (p<.001) and shorter in duration (p<.01 or less). In this context LISA and ROSA were significantly more retracted (lower F2) (p<.001) than NINJA/ASIA. In the segmental possessive context LISA was significantly phonetically higher (lower F1) than NINJA/ASIA (p<.01) (with a similar trend for ROSA). In this context, LISA/ROSA were significantly more retracted (lower F2) than NINJA/ASIA (p<.001). In the syllabic plural context, NINJA was phonetically lower (higher F1) than the other contexts (p<.05 or less), but LISA/ROSA are again significantly phonetically more retracted than NINJA/ASIA with lower F2 (p<.001). LISA/ROSA were also also shorter than NINJA/ASIA (p<.01 or less).

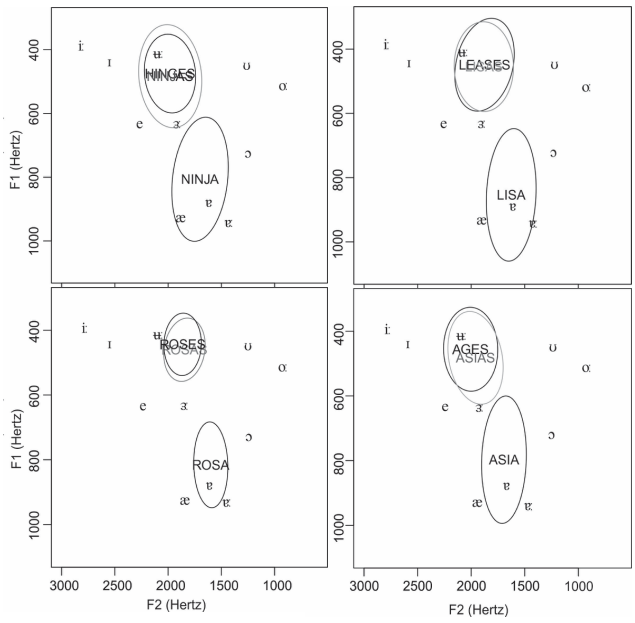


Figure 1: *Position of the unstressed vowel in the F1/F2 plane in the three separate contexts. Ellipses represent two standard deviations from the mean.*

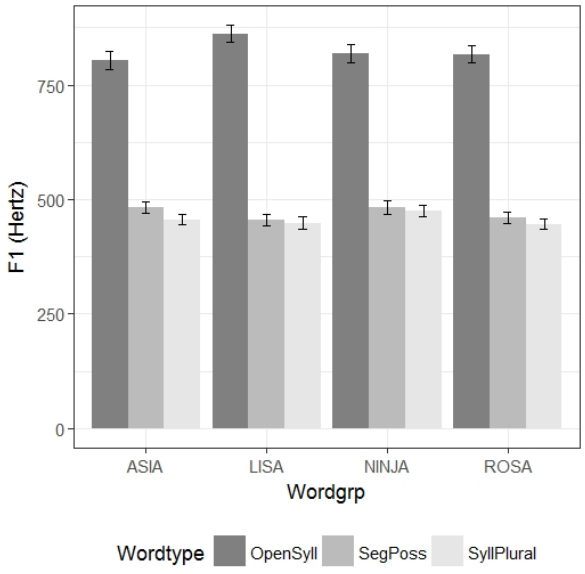


Figure 2: *F1 of the unstressed vowel in hertz across word sets. Whiskers represent confidence intervals.*

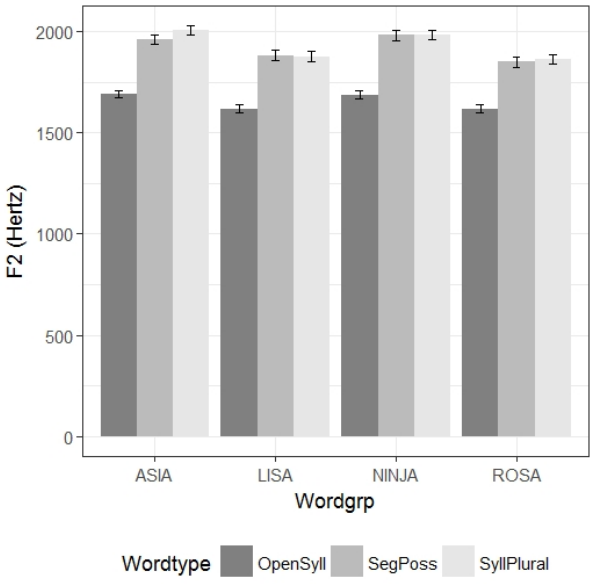


Figure 3: *F2 of the unstressed vowel in hertz across word sets. Whiskers represent confidence intervals.*

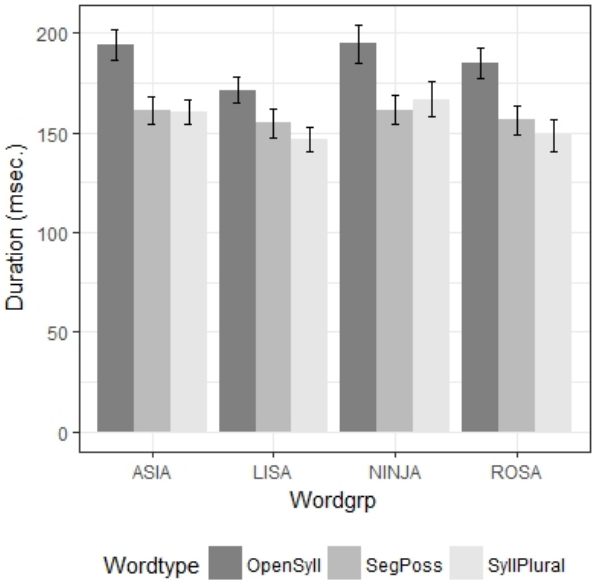


Figure 4: *Duration of the unstressed vowel in milliseconds across word sets. Whiskers represent confidence intervals.*

4. Discussion

The results indicate that schwa in open syllables behaves quite differently from schwa in the preconsonantal context in keeping with the findings of [1] and [16]. This result might be expected as segments before a phrase boundary are typically longer than those that do not have boundary related lengthening ([19], [20]). Phrase final lengthening appears to be progressive. The segment immediately adjacent to the utterance boundary is lengthened the most and the direction of the lengthening is leftward from the right edge of the syllable adjacent to the

utterance boundary [20]. This may explain some of the different durational patterns across schwa types as word-final schwa is adjacent to the boundary. As [9] also suggests, longer vowels have a greater opportunity to reach their designated 'target' which may in turn help to explain the significant formant difference between phrase final schwa and the inter-consonantal schwa. Alternatively, coarticulation is reduced in prosodically strong locations [21], therefore, the open syllable, in addition to being long, is less prone to be affected by coarticulation. The interconsonantal context schwas are flanked by coronals placing considerable coarticulatory pressure on their production as can be seen by the low values for F1. Word-final schwa however, was shown to systematically deviate from the surrounding context suggesting a targetful production.

The significant formant differences across word types can be explained by coarticulation. Each word type comprises a specific phonetic context that affects the schwa uniquely leading to different realisations. In the case of the LISA word type the preceding fricatives is voiceless whereas other word types all have a preceding voiced fricative/affricate. The schwa in word final *Lisa* differs significantly from the other contexts in F1 and duration which may be the result of the preceding voiceless fricative. Vowels flanked by voiceless consonants have a higher F1 [22] although in the interconsonantal context here, when a voiced consonant follows, *Lisa's/leases* does not display raised F1. We note that the word final schwa in *Lisa* was often glottalised perhaps contributing to its shorter duration and raising questions for future analyses. Interestingly, schwa in *Lisa* was the shortest of all the word-final schwas but was also the lowest contradicting previous findings correlating low schwa with length [e.g. 10]. In LISA and ROSA schwa is preceded by alveolar fricatives whereas schwa in ASIA and NINJA follows postalveolar fricatives/affricates. Indeed, the pairs of words with similar preceding fricatives pattern similarly: LISA and ROSA pattern together and ASIA/NINJA pattern together. The only statistically significant difference between the segmental possessive and syllabic plural (affix) occurs for ASIA which can be explained by the variable preceding consonants – *Asia's* /ʒ/ vs *ages* /dʒ/. [23] show that the preceding vowel may also affect the characteristics of schwa further explaining some differences across word types. NINJA also has a different syllabic structure from the other types with a complex preceding consonantal environment in /-ndʒ-/ further complicating the picture.

The height of the possessive and syllabic plural (affix) schwa suggest barred i [ɨ] as a possible allophonic realisation. As only coronal flanking consonants have been included here, it is not possible to say whether this realisation is common for schwa in other non word-final positions. Despite finding a raised vowel in the non word-final coronal context, our results contradict the findings of [16] and [14] that a difference exists between lexical vs affix schwa. We found no significant difference between schwa in segmental possessive and syllabic plural (affix) contexts on any of the dimensions measured: F1, F2, duration (except for the context-specific effect for ASIA – Asia's/ages). Schwa did not display any morphological effect indicating a true Weak Vowel Merger for the items explored in the present dataset of AusE.

5. Conclusions

Results empirically confirm a merger between unstressed vowels in segmental possessive and syllabic plural contexts. Contrary to [14] and [16] we found no effect of morphology on the production of the unstressed vowel for these speakers of AusE. In contrast, word position and segmental context were found to play a major role in the realisation of schwa.

6. Acknowledgements

We would like to thank participants of the Secret 70 Symposium – Festschrift for Prof Andy Butcher 12th October 2017 University of Melbourne for helpful comments on a previous version of this work. We also thank Kelly Miles, Josh Penney, and three anonymous reviewers for insightful suggestions.

7. References

[1] Bernard, J. R. and Lloyd, A. L., "The indeterminate vowel in Sydney and Rockhampton English", in P. Collins & D. Blair [Eds], Australian English: The language of a new society, 288–300, Queensland University Press, 1989.
[2] Wells, J., "Accents of English", Cambridge University Press, 1982.
[3] Trudgill, P., "Dialect contact and new-dialect formation: The inevitability of colonial Englishes", Edinburgh University Press, 2004.
[4] Chambers, J. "Dialect acquisition" Language, 68, 673–705, 1992
[5] Nycz, J. "Second dialect acquisition: A sociophonetic perspective", Language and Linguistics Compass, 9, 469–482, 2015.
[6] Harrington, J., Kleber, F., Reubold, U., Schiel, F. and Stevens, M., "Linking cognitive and social aspects of sound change using agent-based modelling", Topics in Cognitive Science, 1-22, 2018.
[7] Gordon, E., Campbell, L., Hay, J., Maclagan, M., Sudbury, A. and Trudgill, P., "New Zealand English: Its Origins and Evolution", Cambridge University Press, 2004.
[8] Penney, J., Cox, F., and Szakay, A. "Glottalisation of coda stops in Australian English unstressed syllables", Journal of the International Phonetic Association, under review.
[9] Flemming E., "The phonetics of schwa vowels", in D. Minkova [Ed], Phonological Weakness in English, Palgrave. 2007
[10] Keisling, S., "Variation, stance and style", English World-Wide, 26, 1-42, 2005.
[11] Van Bergem, D., "A model of coarticulatory effects on the schwa", Speech Communication, 14, 143-162, 1994.
[12] Fowler C.A., "Production and perception of coarticulation among stressed and unstressed vowels", Journal of Speech and Hearing, 24, 127–139, 1981.
[13] Browman, C. and Goldstein, L., "Targetless schwa: An articulatory analysis", in G. J. Docherty and D. Robert Ladd [Ed], Papers in Laboratory Phonology II: Gesture, Segment, Prosody, 26-56, Cambridge University Press. 1992.
[14] Lammert, A., Goldstein, L., Ramanarayanan, V. and Narayanan, S., "Gestural Control in the English past-tense suffix: An articulatory study using real-time MRI", Phonetica, 71, 229-248, 2014.
[15] Gick, B., "An X-Ray investigation of pharyngeal constriction in American English schwa", Phonetica, 59, 38-48, 2002.
[16] Flemming, E. and Johnson, S., "Rosa's roses: reduced vowels in American English", Journal of the International Phonetic Association, 31, 83-96, 2007.
[17] Kisler,T., Schiel, F. and Sloetjes, H., "Signal processing via webservices: the use case WebMAUS", Digital Humanities, 2012
[18] Boersma, P. and Weenink, D., "Praat: Doing phonetics by computer" Online: http:/www.praat.org/, 2017.
[19] Klatt, D., "Linguistic uses of segmental duration in English: Acoustic and perceptual evidence", JASA, 59, 1208-1221, 1976
[20] Turk, A. E., and Shattuck-Hufnagel, S., "Multiple targets of phrase final lengthening in American English words", Journal of Phonetics 35, 445-472, 2007.
[21] Cho T., "Prosodically conditioned strengthening and vowel-to-vowel coarticulation in English", Journal of Phonetics, 32, 141-176, 2004.
[22] Cole, J., Linebaugh, G., Munson, C. and McMurray, B., "Unmasking the acoustic effects of vowels to vowel coarticulation: A statistical modelling approach", Journal of Phonetics, 38, 167-184, 2010.

Pitch accent variation and realization in interactive discourse in Australian English

Janet Fletcher and Debbie Loakes

School of Languages and Linguistics
University of Melbourne
Centre of Excellence for the Dynamics of Language
janetf;dloakes@unimelb.edu.au

Abstract

This paper examines pre-nuclear and nuclear pitch accent variation and realization in map task interactions for Standard Australian English. An analysis of pitch accent choice in accented words reveals that pitch accent type is not always critical to the realization of information structure categories; for instance in the case of informational focus participants often use a range of pitch accent shapes. Nevertheless nuclear bitonal accents (e.g. L+H* pitch accents) are scaled higher than simple H* pitch accents across the board in line with previous findings for other English varieties. Furthermore there is evidence of an interaction between dialog act, pitch accent choice and realization with more rising and/or higher scaled accents used in particular kinds of interactions.

1. Introduction

The majority of research on Australian English intonation has focused on the prevalence and functions of Uptalk in interactive discourse [see 1 for a comprehensive overview]. Uptalk refers to the use of rising (often high rising) pitch at the end of syntactic declarative statements. Conversely in previous studies of Australian English intonation there has been relatively little focus on post-lexical pitch accent variation, particularly with regard to pitch accent type and pitch accent scaling, compared to other varieties of English. Earlier corpus studies for British English varieties [2] and experimental studies of American English [e.g. 3] have revealed that there is strong evidence of variation in terms of pitch accent type and pitch accent realization across different speech communities. Moreover, it has generally been assumed for Australian English that, as in Southern British English and General American English, pitch accent choice and realization play a crucial role in signaling information structure categories (e.g. focus) in spoken communication. With regard to the realization of neutral, informational, and contrastive focus, studies of American English have shown that a high target pitch accent with a strong lead tone, e.g. L+H*, tends to be used in narrow and contrastive focus contexts rather than a simple H* pitch accent [3, 4] although there is some evidence of variation among different American varieties [e.g. 3]. Moreover the H* tone target of the rising L+H* accent is generally scaled higher in a speaker's range in these narrow and contrastive focus contexts. Similar effects of focus have been found in many different languages [e.g. see 5, 6, 7, 8 for an overview]. Preliminary acoustic studies of Australian English also suggest that rising accents play an important role in signaling contrastive focus in carefully scripted laboratory phonology type experiments [e.g. 9].

An example of a potential contrastive focus context is shown below. In this example (utterance b), *MANLY* would potentially be realized with a rising L+H* rather than simple H* accent (both shown schematically in Figure 1) to emphasise that the interlocutor went to Manly and not Bondi. The simple H* shown on the left does not have the strong rising lead tone that characterizes the L+H* pitch accent on the right (located on the main stressed syllable of *Manly*). According to the standard ToBI descriptions of these accents for other varieties of English, the H* target of the bitonal L+H* accent is also likely to be scaled higher in the speaker's pitch range as shown in Figure 1.

a) Did you go to Bondi baths yesterday?
b) No I went to MANLY baths.

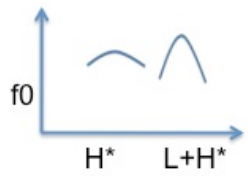


Figure 1. Schematic representation of two major pitch accent f0 shapes: H and L+H*.*

It remains to be seen whether rising pitch accents are widely used in less scripted data in Australian English in general, and/or whether they are used frequently to signal contrastive focus in spoken interaction compared to what has been observed in controlled laboratory-style speech. It also remains to be seen whether pitch accent variation constitutes another source of potential differentiation between Standard Australian English and other so called "mainstream" varieties of English. It may be that rising accents, and not only high-rising boundaries, contribute to the general perception of Australian English as a 'rising' variety but this remains to be quantitatively verified, e.g. [1].

The main focus of this study is to investigate pitch accent variation and realization in a corpus of map task data (Australian National Database for Spoken Language, henceforth ANDOSL) for Australian English [10]. The additional aim of this study is to provide a "reference" dataset for future comparison with the AusTalk data collection which also includes a map task [11]. It should be noted that the ANDOSL Map Task used here was collected in Sydney in the '90's and is indicative of Standard Australian English spoken in the Sydney/NSW region at that time, although it should also be noted there is no indication that prosody has changed over time in Australian English. This study will give a further

understanding of (rising) prosodic patterns in spontaneous Australian English.

1.1 ANDOSL Map Task

The Map Task is an exercise in controlled quasi-spontaneous talk between two participants and therefore constitutes a rich resource for the investigation of situated interaction and in particular the interaction between intonational categories like pitch accents and information structure (analysed in the current study). The structure of a map task can be summarized as follows: participants work in pairs, each with a map in front of them that the other participant cannot see. The maps contain a number of landmarks, which are not all identical. One participant (the 'instruction-giver' IG) has a route marked on his/her map and is required to instruct the other participant (the 'instruction-follower' IF) to draw the correct route onto their own map. The mismatching landmarks on each map ensure that a range of queries, checks and negotiation talk will be elicited. This is one reason why map tasks are an effective tool to examine intonational variation given the known relationship of intonation to utterance modality and discourse segmentation (e.g. [12]).

2. Method and Materials

2.1 Participants

Eight dialogs (4 speaker pairs) from the Australian Map Task corpus of the Australian National Database [10] formed the dataset for this study. This was a subset of the corpus that was analyzed in earlier studies of Uptalk and high rising tunes in Australian English (e.g. [13]) and studies of discourse modeling (e.g.[14]). Each speaker pair consisted of a male and female who took it in turns to be Instruction Giver (IG) or Instruction follower (IF). The dialogs were chosen randomly and the speakers all belong to the standard Australian English dialectal grouping. The dialogs were between 485.93 sec and 810.24 sec in duration and were digitized at 22,500 kHz.

2.2 Word and prosodic labeling

The map data were annotated according to ToBI (Tones and Break Indices) conventions that have been adapted for Australian English ([15]). Fundamental frequency contours (f0) were annotated for pitch accents. Major pitch targets or movements corresponding to pitch accents (e.g. H*, L+H) intermediate phrase boundaries (e.g. L-H-) and intonational phrase boundaries (e.g. L% H%) were labelled using the f0 signal and auditory analysis by four expert ToBI transcribers. There were 2831 intermediate or intonational phrases in the dataset with the equivalent number of nuclear accents (recall that a nuclear accent is the head of an intermediate phrase in the AM model that underpins AusE_ToBI), and 1964 pre-nuclear accents (7626 tokens in total). Word boundaries were manually identified from speech waveforms and spectrograms and orthographically annotated using *ESPS/xwaves* and *emu-labeller* [16]. Pitch accent categories were chosen from the main set of tonal categories proposed for AusE_ToBI: L* (low tone target realized in the lowest part of a speaker's pitch range), H* (high tone target corresponding to a shallow peak in the higher part of a speaker's pitch range), L+H* (high tone target with strong rising lead tone), L*+H tone target (low tone target on accented syllable with late rising tone), and downstepped variants including !H* (lowered pitch peak relative to a pre-nuclear accent in the same intermediate phrase).

Of particular interest in this study is variation in simple (H*, !H*) versus bitonal (L+H*, L*+H) pitch accent choice, and scaling of the L and H targets in pre-nuclear and nuclear contexts. An example of intonational annotation illustrating two different pitch accent types, H* and L+H* respectively is shown in Figures 2 and 3. Both are examples of contrastive focus constructions which were of additional interest in our study. Due to the mismatch between the two maps used by the participants in the task, the IFs often produced these kinds of constructions in response to a request by the IG. The illustration in Figure 2 shows the second part of an exchange between an IG and an IP pair which is shown below.

IG *"Do you have a galah open-cut mine?"*
IF *"I've got a DINGO."*

The IF responds to an information question posed by the IG about whether or not he has a particular landmark on his map, in this case "the galah open cut mine". The IF has a "mine" with a different name on his map and so in his response, the word "DINGO" is realized with a simple H* nuclear accent scaled in the upper part of this speakers' pitch range, rather than a L+H* bitonal accent (there is no strong lead tone evident on the accented first syllable of "Dingo").

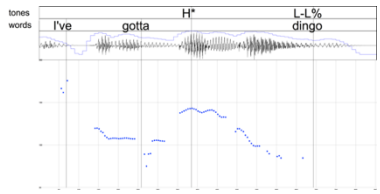


Figure 2. F0 contour showing an example of a nuclear H pitch accent produced by a male participant (Instruction follower) on the accented first syllable of "Dingo" in a map task dialog.*

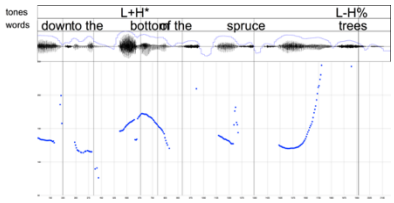


Figure 3. F0 contour showing an example of a nuclear L+H pitch accent produced by a male participant (Instruction Giver) on the accented first syllable of "bottom" in a map task dialog.*

Figure 3 by contrast shows a L+H* pitch accent on the nuclear accented word "bottom" in the utterance "down to the *bottom* of the spruce trees". In both cases the boundary tones also differ. This kind of pitch accent variation suggests that we may find more variation in these kinds of quasi-spontaneous interactions than has been found in more scripted laboratory tasks.

2.3. Data Analysis

f0 was calculated using the Schaeffer-Vincent algorithm in the *Wrassp* library in *emuR* [18] and values extracted for pitch accent categories at the location of the annotated starred tone. For H*, !H* and L+H* pitch accents this was usually late in the rhyme of the accented syllable and for L* and L*+H accents this was located in the middle of the (low) pitch elbow

associated with the accented syllable rhyme. The f0 values extracted at the * tone target were converted to semitones (benchmark 50). A linear mixed effects analysis was then performed in R (*lmerTest* and *step* [19]) to compare pitch accent scaling across the corpus. A maximally specified mixed model was implemented with fixed factors and interactions for PITCH ACCENT, SEX, POSITION (nuclear versus pre-nuclear) and map task ROLE (Instruction Giver [IG], Instruction Follower [IF]) with random factor SPEAKER. Post-hoc comparisons using Bonferroni correction were undertaken to investigate any significant interactions. It is well-attested that females have higher mean pitch than male speakers and it was assumed that this would be a significant factor. The main point of this analysis was to explore any simple main effects and potential interaction between speaker role and pitch accent scaling differences, with a particular focus on rising versus simple high pitch accents in nuclear versus pre-nuclear contexts.

3. Results

3.1. Nuclear and pre-nuclear pitch accent distribution

The distribution of major simple and bitonal accents in the eight maptasks is shown in Figure 4 for speakers in the IG role, and in Figure 5 for speakers taking the IF role.

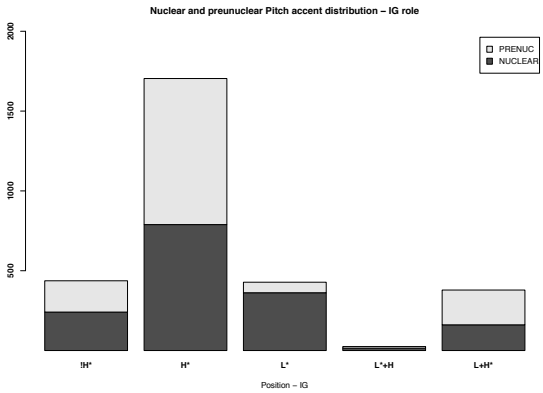


Figure 4. Distribution of simple and bitonal (two tone) pitch accents in nuclear (black) and pre-nuclear (grey) position - Instruction Giver (IG) role.

There is similarity in nuclear pitch accent distribution across the two roles with some exceptions. Simple H* accents are the most frequent accent-type across the corpus (more than 50% of all accents produced by IGs, and 62% of all accents produced by IFs). L+H* pitch accents also significantly outnumber the instances of L*+H pitch accents. The higher proportion of simple high accents in the Instruction Follower sections of the task can be related to the lower number of downstepped !H* accents. Speaker turns and utterances tended to be shorter in IF contexts [see 17] with a high number of IPs and dialog acts consisting of a single pitch accent and boundary tone configuration. With regard to rising (L+H*, L*+H) versus simple H* or !H* accents, around 13% of nuclear accents are rising in this corpus. The effect of speaker role (IG or IF) was not significant (χ^2=3.3351, $p > 0.05$) in contrast to speaker SEX which was (χ^2=9.3775, $p < 0.001$). Females produced more bi-tonal rising accents (L+H*) in nuclear position than males who also produced fewer L*+H accents than females. Pre-nuclear accents show a broadly similar distribution - 60% of pitch accents are simple H* and around 15% are rising L+H* accents and there was no significant effect of speaker SEX ($p > 0.05$) in terms of pitch accent selection.

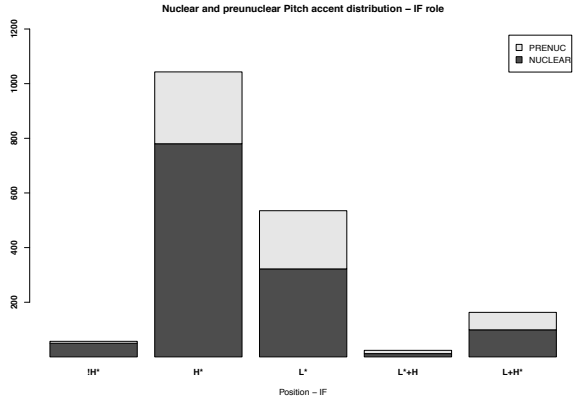


Figure 5. Distribution of simple and bitonal (two tone) pitch accents in nuclear (black) and pre-nuclear (grey) position - Instruction Follower (IF) role.

3.2. Pitch scaling of nuclear and pre-nuclear accents

Pitch level (scaling) values in semitones (ST) for the different pitch accent types (simple and bitonal) in nuclear position are shown in Figure 6 for female participants and in Figure 7 for male participants in the Instruction Giver (IG) and Instruction Follower (IF) roles. Notwithstanding the predictable effect of speaker SEX (all pitch accents are scaled higher for females compared to males), there were statistically significant main effects for ROLE, POSITION (nuclear versus pre-nuclear) and PITCH ACCENT, with interactions between POSITION and PITCH ACCENT and ACCENT and ROLE.

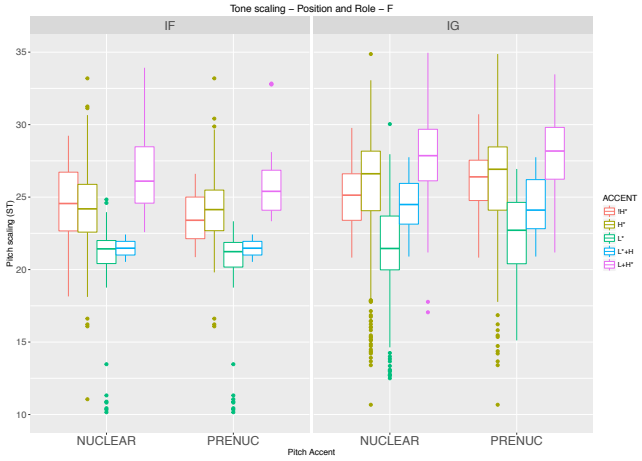


*Figure 6. Pitch level (semitones) of the *tone of pitch accents produced by female participants in the Instruction follower role (left panel) and instruction Giver role (right panel*

Post-hoc tests (Bonferroni) confirm that the differences in scaling between rising accents (L+H*) and simple H* accents is significant in both speaker roles for males and females (M: $t=14.73$, $p<0.0001$; $t=-9.95$, $p<0.0001$). It should also be noted that there is a degree of overlap between pitch values for H* pitch accents and the H tone of L+H* pitch accents. In terms of speaker ROLE, H* accents are scaled higher for IGs versus IFs ($t=-5.78$, $p<0.0001$), but there is no significant scaling difference for L+H* accents according to ROLE. Nuclear H*

and L+H* accents are scaled higher than pre-nuclear high tone accents (H*: t=-5.77, p<0.0001; L+H*: t=-13.41, p<0.0001) in both roles with nuclear L+H* accents consistently scaled higher than H* accents. Pre-nuclear !H* or L* accents are not scaled higher or lower than their nuclear counterparts.

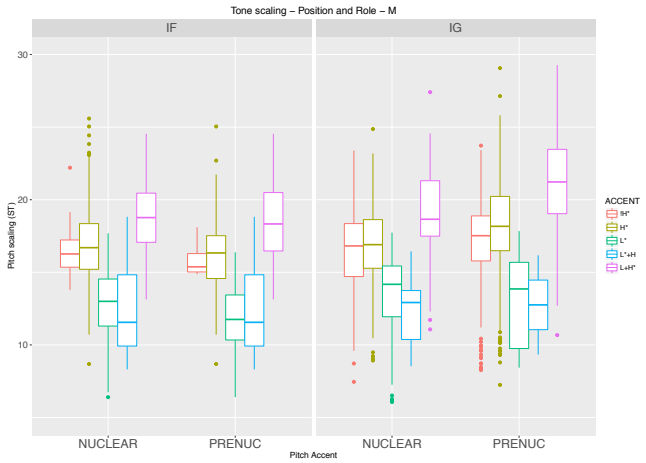


*Figure 7. Pitch level (semitones) of the *tone of pitch accents produced by male speakers in the Instruction follower role (left panel) and instruction Giver role (right panel).*

4. Discussion

Australian English speakers use a variety of pitch accents in these map task interactions. By far the most common accent type is the simple H* accent which reflects a pattern observed in corpus studies of other varieties including Mainstream American English, although proportionally fewer rising accents are actually observed in our corpus compared to [20]. This suggests that the observed 'rising' nature of Australian English is actually due to the high incidence of intonational phrase-final rises, i.e. Uptalk [see 1, 17], rather than frequency of rising pitch accents within phrases. The contrastive status of the L+H* versus H* accent has also often been a source of debate in intonational phonology [e.g. 21] and it could be that transcribers were conservative in their use of this accent category in this corpus given its well-known ambiguity with H* accents [20]. Nevertheless, the H tone of transcribed nuclear and pre-nuclear rising accents (L+H*) is consistently scaled higher than simple H* accents in these Australian English map task interactions, in keeping with earlier findings for other English varieties [3, 20], although there is unsurprisingly a degree of overlap observed in accent realizations.

These findings also support observations from scripted laboratory-style speech for younger Australian English speakers from Melbourne [9], although in this study (using unscripted speech) it is not always the case that rising accents are used exclusively in contrastive focus or informational focus constructions. Higher scaled H* accents are more commonly used, as well as the occasional use of H+!H*. Rising accents also tend to be used in informational focus contexts where one participant wishes to draw the other participant's attention to pertinent information in relation to the task. Rising accents also occur in specific types of dialog acts including action directives to influence the action of the addressee i.e. the IF, and responses to the IG. A preliminary analysis of dialog acts which include contrastive focus contexts (e.g. cases of negative responses to instructions by the IG), suggests that speakers produce both single high tone and rising pitch accents in these focus contexts, although both accent types are typically scaled higher than pre-nuclear accents in the same intonational/intermediate phrase. This suggests that the normal downtrend patterns often observed in an intonational phrase are not apparent in these cases. Female participants also tend to use more rising accents than male participants. It may be that the expanded pitch range associated with female speech makes it easier for transcribers to annotate a rising versus simple high pitch accent, or we may be tapping into pragmatic and/or sociophonetic variation amongst Australian English speakers, but this requires further investigation. In summary, intonational analysis of situated interaction (even in quasi-spontaneous speech tasks) produces significant patterns of variation that should be captured by our intonational models, and gives a greater understanding of the nuanced ways speakers use prosody in spoken communication.

5. References

[1] Warren, P. *Uptalk: The Phenomenon of Rising Intonation.* Cambridge University Press, 2016.

[2] Grabe, E. et al. "Pitch accent realization in four varieties of British English", *Journal of Phonetics*, 27, pp.161-185, 2000.

[3] Garding, G. and Arvaniti, A. "Dialectal Variation in the Rising Accents of American English", *Papers in Laboratory Phonology IX: Change in Phonology*, pp. 547-576, 2007

[4] Liberman, M. & Pierrehumbert, J. "Intonational invariance under changes in pitch range and length". In M. Aronoff & R. Oehrle (eds.), *Language Sound Structure* MIT Press, pp. 157-233, 1984.

[5] Ladd, D.R. *Intonational Phonology*, CUP, 2008.

[6] Gussenhoven, C. *Phonology of Tone and Intonation.* CUP, 2004.

[7] Féry, C. *Intonation and Prosodic Structure*, CUP, 2016.

[8] Baumann, S. & Kügler, F. "Prosody and information status in typological perspective", *Lingua*, 165, pp. 179-182, 2015.

[9] Holt, C., & Fletcher, J. "Pitch accentuation in adolescent users of cochlear implants", *SST2010*, pp. 138-141, 2010.

[10] Millar, B et al. "The Australian National Database of Spoken Language", *Proceedings of ICASSP-94*, pp. 197-100, 1994.

[11] Estival, D. "AusTalk and Alveo: An Australian Corpus and Human Communication Science Collaboration Down Under". In *Language Production, Cognition, and the Lexicon*. Núria Gala et al. eds, Vol. 48. Springer. pp. 545-560, 2015.

[12] Hirschberg, J. "Communication and prosody: Functional aspects of prosody. *Speech Communication*, 36, pp31-43, 2002.

[13] Fletcher J. & Stirling, L. "Prosody and Discourse in the Australian Map Task Corpus.", In J. Durand et al. (eds.), *The Oxford Handbook of Corpus Phonology*, OUP, 562-575, 2014.

[14] Mushin, I.et al. "Discourse structure, grounding, and prosody in task-oriented dialog", *Discourse Proc.*, 35 (1), pp. 1-32, 2003.

[15] Fletcher, J. & Harrington, J. "High-rising terminals and fall-rises in Australian English", *Phonetica*, 58 (4), 215-220, 2001.

[16] Winkelmann, R., & Raess, G. "Introducing a Web Application for Labeling, Visualizing Speech and Correcting Derived Speech Signals". *LREC'14*, Reykjavik, Iceland, pp 4129-4133, 2014.

[17] Fletcher, J., Stirling, L., Wales, R., & Mushin, I. "Rising intonation and dialog acts in Australian English", *Language and Speech*, 45 (3), pp. 229-254, 2002.

[18] Winkelmann, R. et al., emuR version 0.2.2. Available from: https://github.com/IPS-LMU/emuR, 2017.

[19] KuznetsovaA."Packagelmertest". centos.ustc.edu.cn/CRAN/web/packages/lmerTest/lmerTest.pdf 2016

[20] Brugos, A., N. Veilleux, M. Breen & S. Shattuck-Hufnagel "The alternatives (ALT) tier for ToBI: Advantages of Capturing Prosodic Ambiguity", *Speech Prosody 2008*, http://aune.lpl.univ-aix. fr/sprosig/sp2008/papers/id072.pdf

[21] Ladd, D.R. & Schepman, A. "Sagging transitions between high accents in English: Experimental evidence", *Journal of Phonetics*, 31 pp 81-112, 2003.

Gender differences in the spectral characteristics of voiceless sibilants produced by Australian English-speaking children

Casey Ford, Marija Tabain* & Gerry Docherty^*

*La Trobe University, Melbourne, Australia
^Griffith University, Brisbane, Australia
ceford@students.latrobe.edu.au, m.tabain@latrobe.edu.au, gerry.docherty@griffith.edu.au

Abstract

This paper examines the spectral characteristics of the voiceless sibilant fricatives /s/ and /ʃ/ produced by Australian English-speaking children (5-13 years) from a town in rural Victoria. It finds that despite the lack of sex dimorphism in the vocal tract, gender differences are evident in sibilant production. Girls produce sibilants with higher spectral mean and lower spectral skewness than boys, even for the youngest speakers examined. Spectral changes over time suggest potential influence from social factors such as region and socioeconomic status. Results provide a basis for the development of sociophonetic variation of sibilants and Australian English more broadly.

Index Terms: Australian English, fricatives, spectral moments, sociophonetics, language acquisition.

1. Introduction

Sociophonetic studies on the production of voiceless fricatives, particularly voiceless sibilants /s/ and /ʃ/, have shown that certain spectral features correlate with particular social factors. Sex and gender are particularly robust social factors in sibilant production. Studies that have focused on sibilants have provided much evidence of clear male-female differences in their acoustic characteristics, most commonly among adult speakers of English. The origins of these acoustic differences are held partly in physiology, where sex dimorphism in vocal tract morphology, the size and shape of the constriction in the production of the sibilant, the position of the tongue tip and teeth, and lip rounding all affect the spectral characteristics of the fricative [1]. It is expected that females are more likely to have smaller vocal tracts than males, and it is therefore assumed that the resulting production of sibilants for female speakers will have higher corresponding spectral frequencies than males [2].

While physiology may go some way in explaining sex differences in sibilant characteristics, there is much evidence to suggest that social and behavioural factors are more likely sources of this variation. Adjusting the sibilant point of constriction can affect the resulting resonances: a retracted point of constriction gives rise to lower frequencies, while a more fronted articulation results in higher frequencies. An expanding body of work has given much support for such articulatory behaviours to be socially influenced, and acoustic data has shown that speakers adjust their articulation within their physiological limits in order to align themselves with a particular social identity. The production of higher frequency sibilants as a result of a fronted articulation is perceived as a characteristically feminine and stereotypical of women and gay men [3], [4], while masculinity and working class associations are drawn from lower frequency sibilants produced with a retracted articulation [5], [6]. Regional affiliation has also been shown to influence sibilant production, where rural speakers appear to produce more retracted sibilants in comparison to their urban counterparts [7].

Examining the acoustic characteristics of children's speech particularly highlights the importance of social and cultural influences on phonetic variation. It is generally acknowledged that prepubescent children exhibit little to no sex-based dimorphism in vocal tract morphology [8]. Despite this, there has been consistent evidence of clear gender differences in the acoustic measurement of children's sibilants, where spectral characteristics correlate with those expected for their gender [9], [10]. Girls are reported to produce sibilants with higher spectral frequencies than boys of the same age. These differences become larger with an increase in age due to the onset of structural changes in the vocal tract for male speakers at puberty. Other work has also shown that gender-specific variation of plosives and connect speech processes is evident for young speakers prior to adolescence [11], [12]. These gender-specific features in their speech have been attributed to articulatory behaviours children learn from their social environment.

Little is known about sociophonetic variation of sibilants in Australian English in general. Therefore, this paper aims to address two gaps in the current literature. First, it examines gender-specific production of sibilants by children from early childhood to preadolescence (5-13 years). It also aims to give a basis for the development of sociophonetic variation specific to the Australian English context. This study will measure sibilant spectral moments to examine the effects of gender and age over the primary school years. It is hypothesised that speakers in all age groups will produce sibilants with spectral characteristics that align with those expected for their gender, and that gender differences will become greater with an increase in age.

2. Method

2.1. Speech community, participants, and recording

Data were collected by the first author as part of a larger PhD project. The children attended a government-funded primary school in the rural town of Yarrawonga, Victoria, located approximately 270km north-east of Melbourne, with a population of around 8000. It is a town of relatively limited cultural diversity, where the majority of its citizens are of an Anglo Celtic background and identify as 'Australian' in ancestry. Yarrawonga is primarily a monolingual town, and only around 4% of households speak a language other than

English. Its major industries are in the manufacturing, construction, and tourism and hospitality sectors. It is classified as a town of relatively high socioeconomic disadvantage [13]. Data were collected here as a starting point for examining sociophonetic variation between rural and urban speakers of Australian English. 46 speakers were recruited from three primary school year levels (henceforth referred to as age groups): Prep, Year Three, and Year Six. These age groups were selected in order to facilitate an overview of the seven-year primary school period – a crucial time in the acquisition of sociolinguistic competence [14]. Here, data are from 34 of the 46 speakers recorded. Information about each speaker group is reported in Table 1.

Speakers were recorded in a quiet room on the school campus during school hours in sex- and age group-matched dyads. They participated in a casual, interactive *sociolinguistic interaction* which involved a combination of laboratory-style tasks and more traditional sociolinguistic interview-style tasks. All recordings were made using a Marantz Professional PMD661 solid-state recorder and two Shure SM94 microphones at a sampling rate of 44.1 kHz. Each speaker had a separate microphone placed in front of them on a boom stand. Microphones were placed at around 20-30cm in front of the speaker's mouth, or as close to this as possible.

Age group	Sex	No.	Age range (y;m)	Mean age (y;m)	/s/ tokens	/ʃ/ tokens
Prep	Girls	6	5;7-6;2	5;11	441	71
	Boys	6	5;4-6;4	5;9	646	96
Year Three	Girls	6	8;3-9;3	8;8	852	94
	Boys	6	8;2-9;6	8;10	1127	99
Year Six	Girls	6	11;8-13;0	12;4	1351	169
	Boys	4	11;10-12;8	12;3	1070	127

Table 1 *Speaker group information.*

2.2. Analysis

2.2.1. Data preparation

Each speaker's recordings were broken down into short, manageable sections, transcribed, and then segmented and force-aligned using the WebMAUS-multiple automated alignment service [15]. Segment boundaries were manually adjusted using the *Emu Speech Database System* [16]. Fricatives were segmented by placing the onset boundary at the onset of the frication noise, and then offset boundary at the cessation of the frication noise and the onset of the following segment. Any tokens adjacent to another fricative sound, or tokens that contained instances of interference, such as interruption from another speaker or other background noises, were omitted. Across the 34 speakers, 6213 tokens were extracted. 69 of these tokens were omitted, leaving 6144 tokens for analysis. 5488 of these were tokens of /s/, and the remaining 656 were /ʃ/. Tokens were extracted using the Emu/R interface [17]. Token numbers for each speaker group are shown in Table 1. Prosodic context, surrounding vowel contexts, and effect of speech task were not considered for the current study, but will be examined in future work.

2.2.2. Spectral moments

Fricative spectra were estimated using Fast Fourier Transform (FFT) with a 20ms Hamming window over the fricative midpoint. Spectral moments [18] were extracted at the temporal midpoint of each fricative within a spectral range of 1 – 15 kHz in order to capture the majority of the energy distribution while filtering out any potential background noise or adjacent voicing. The midpoint was measured to avoid the influence of co-articulatory gestures at the fricative onset and offset and any surrounding vowel formant transitions, making it the ideal place to examine the spectral qualities of the sibilant itself.

The measurement of spectral moments is a common and robust approach in examining sociophonetic influences on fricative production. Typically, the first four spectral moments are measured, however this paper will focus only on the first (M1) and third (M3) spectral moments. These have been shown to be particularly sensitive to speaker sex [3], [9]. M1 represents the spectral mean and reflects the average energy concentration. It gives the midpoint frequency at which the distribution of energy on either side of it is equal. M3 gives a measure of skewness by measuring the energy above and below the mean and indicates asymmetry in the energy distribution. A positive skewness value indicates the concentration of energy is in the lower frequencies, while a negative skewness value indicates the opposite. In relation to sibilant articulation, a more fronted articulation will not only give a higher peak frequency, but also higher M1 and lower M3 values.

2.2.3. Statistical analysis

Linear mixed effects models were built to test the statistical effects of the social and linguistic factors on the spectral properties of the sibilants. The spectral parameters (M1 and M3) served as dependent factors. Effects of sex, age group, and sibilant, and sex × age group × sibilant interactions were set as independent fixed factors. Speaker was included as a random intercept. Sibilant was set as a within-subject factor. This was carried out using the *lmerTest* [19] package in R, with degrees of freedom estimated using the Satterthwaite approximation method. Post-hoc Tukey pairwise comparisons were carried out using *lsmeans* [20]. Significance was set at 0.05.

3. Results

3.1. First spectral moment – spectral mean (M1)

Figure 1 shows the average M1 values for the two sibilants for each sex in each of the three age groups. It shows that girls' average M1 is higher than the boys' for both sibilants in all three age groups. This gender difference is consistent with what has been found in previous work. Across the age groups, the boys show a decrease in M1 for both sibilants with an increase in age, with the main drop in frequency occurring between Year Three and Year Six. The girls' M1 values for /s/ change in parallel with the boys'. A decrease with an increase in age is apparent, with the main drop in frequency occurring between the two older age groups. Their /ʃ/ M1 also decreases over time, but the main decrease occurs between Prep and Year Three. The girls' decrease in M1 with an increase in age is interesting, as previous work has shown girls' M1 values to stay relatively steady during this period [9]. The boys' decrease in M1 over time is expected, especially given the possibility for 12-year-old boys to be pubertal.

Statistically, speaker sex, age group, and sibilant all emerged as significant effects on M1, as shown in Table 2. However, none of the interactions emerged as significant. Pairwise comparisons were undertaken to investigate the patterns within the data, the results of which are shown in

Table 3. Only significant results (p <.05) are reported in the interest of brevity. For speakers overall, the girls' /s/ and /ʃ/ M1 was significantly higher than the boys'. The decrease in the boys' /s/ with an increase in age was also significant, however this was not the case for their /ʃ/. The decrease in the girls' M1 with an increase in age was also not significant for either sibilant.

While girls' /s/ M1 is higher than boys at all age groups, statistically significant differences were not apparent until Year Six. For /ʃ/, M1 was significantly higher for the girls in the youngest age group compared to the boys, however sex differences in the two older age groups were not significant.

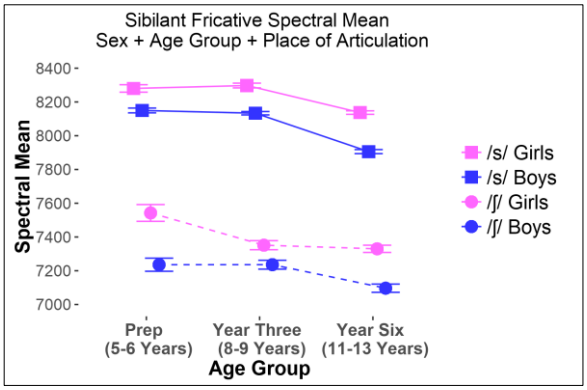


Figure 1 *Spectral mean of the two sibilants by speaker sex and age group. Error bars indicate standard error.*

Effects	Model output	*p*-value
sex	F[1,29.28]=13.63	<.001
age group	F[2,29.23]=5.06	<.05
sibilant	F[1,32.07]=541.89	<.0001

Table 2 *Significant effects on M1.*

Factors	Comparison	Difference (Hz)	Std. Error	Df	t-ratio	*p*-value
Sibilant between sexes	Girls – Boys /s/	194	61.5	41	3.16	<.05
	Girls – Boys /ʃ/	204	67.8	60	3.01	<.05
sibilant *within* sexes *between* age groups	Prep – Year Six Boys /s/	299	114.5	42	2.61	<.05
	Year Three – Year Six Boys /s/	296	113.7	41	2.61	<.05
sibilant *between* sexes *within* age groups	Year Six Girls – Boys /s/	315	113.7	41	1.29	<.01
	Prep Girls – Boys /ʃ/	266	116.7	68	2.28	<.05

Table 3 *Significant pairwise comparisons for M1.*

3.2. Third spectral moment – spectral skewness (M3)

Figure 2 shows the average M3 values of the two sibilants by speaker sex and age group. Boys' M3 values for both sibilants are higher than girls' in each of the three age groups, indicating a concentration of fricative energy in lower frequencies for the boys. Across the age groups, boys show an increase in M3 for both sibilants with an increase in age. This increase is consistent across age groups for /ʃ/, while the main increase for /s/ occurs between Year Three and Year Six. The girls' M3 also increases for both sibilants with an increase in age. For /s/, the main change in skewness occurs between Year Three and Year Six, and between Prep and Year Three for /ʃ/. These patterns mirror those found in their M1.

Statistically, sex, age group, and sibilant all emerged as significant main effects on M3, as seen in Table 4. Again, none of the factor interactions emerged as significant. Pairwise comparisons, shown in Table 5, provide more information about the patterns found for M3. The girls' M3 values are significantly lower for /s/ but not for /ʃ/. Between age groups, the increase in skewness for the boys' /s/ is significant. As for sex differences in specific age groups, these become greater with an increase in age. Significant differences are apparent in Year Three and Year Six, where the girls' M3 values are significantly lower than the boys' in both age groups.

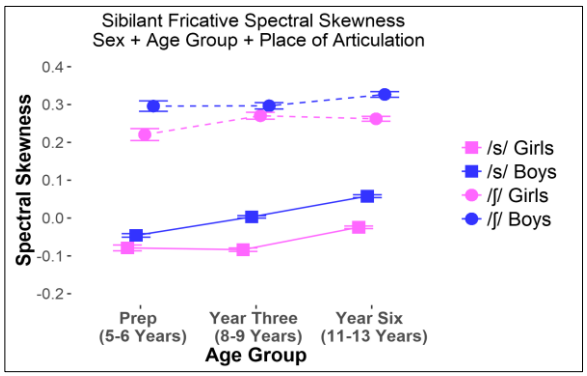


Figure 2 *Spectral skewness of the two sibilants by speaker sex and age group. Error bars indicate standard error.*

Effects	Model output	*p*-value
sex	F[1,29.00]=12.73	<.01
age group	F[2,28.95]=4.36.	<.05
sibilant	F[1,33.47]=605.87	<.0001

Table 4 *Significant effects on M3.*

Factors	Comparison	Difference	Std. Error	Df	t-ratio	*p*-value
sibilant between sexes	Girls – Boys /s/	-0.07	0.02	42	-3.51	<.01
sibilant *within* sexes *between* age groups	Prep – Year Six Boys /s/	-0.11	0.04	42	-2.86	<.05
sibilant *between* sexes *within* age groups	Year Three Girls – Boys /s/	-0.07	0.03	42	-2.13	<.05
	Year Six Girls – Boys /s/	-0.10	0.04	41	-2.72	<.01

Table 5 *Significant pairwise comparisons for M3.*

4. Discussion

Gender differences in sibilant production are evident for these prepubescent speakers of Australian English. In all three age groups examined, girls produced both sibilants with higher spectral mean and lower spectral skewness compared to boys. These results are consistent with previous findings on gender-specific sibilant spectral characteristics as produced by adult speakers [3], and previous findings on gender differences in children's sibilants [9]. It is important to note that these differences are apparent for speakers as young as 5 years of age. Significant differences in some spectral features begin to appear at around 8 years of age (Year Three), but differences are most significant at around 12 years of age. Gender

differences in sibilant production were most significant and consistent for /s/. This result is not surprising, given the socioindexical power of the voiceless alveolar fricative.

Given the unlikely leading role of physiology in the gender differences for speakers examined here, behavioural and social factors appear to be a major influence on sibilant articulation and the resulting spectral characteristics. The results found in the changes in spectral frequencies with an increase in age for boys are consistent with those found in previous work, where lowering with age is expected [9]. It is possible for the oldest speakers in this study to be at the beginning of puberty, leading to a lowering of frequency characteristics in the male voices. However, the parallel decrease in the girls' sibilants was an unexpected result. Although some decrease might be expected due to an increase in size of the oral cavity with growth, previous work with similarly aged girls showed their sibilant spectral frequencies to stay relatively steady [9]. A similar result was found in the voiceless plosives produced by a subset of the speakers examined here, where the girls used fewer female-oriented variants and more male-oriented variants of /p, t, k/ with an increase in age [11].

The intersection between gender and other social factors such as rurality or socioeconomic status may offer some explanation for the pattern exhibited in the girls' sibilants as they approach adolescence, particularly for /s/. Work on /s/ production in rural and urban areas of California [7] has found that rural women produce /s/ with lower spectral frequency than women in urban centres. Moreover, women in rural towns who have more affinity with the country and rural life produce /s/ with lower spectral frequencies than rural women who have more of an interest in city life. In terms of socioeconomic status, work on /s/ in Glaswegian English, for example, found the young working class women produce /s/ with articulatory and acoustic features more similar to men than to young middle class women [5]. In an Australian context, there is an inherent connection between working class, rurality, and ideals of masculinity [21]. While Year Six girls still produce /s/ with significantly higher spectral mean than the boys, growing up in rural Australia and conforming with peer group norms may influence girls' lowering of /s/ spectral frequencies, and increased use of more male-related features.

It is clear from the current results that sociophonetic variation in Australian English sibilants begins at a young age, particularly in relation to gender. In order to learn more about sociophonetic features of Australian English, and to corroborate the patterns found here, comparing sibilants produced by adults from the same speech community will give further insight into sociophonetic acquisition in Australian English. Moreover, comparing these patterns with sibilants produced by children and adult speakers in urban areas will give further insight into sociophonetic variation in Australian English more broadly.

5. Acknowledgments

Many thanks go to the staff and students at YCP-12, and to the anonymous reviewers for their helpful comments on an earlier version of this paper. All remaining errors are the responsibility of the first author.

6. References

[1] C. Shadle, "Articulatory-acoustic relationships in fricative consonants," in *Speech Production and Speech Modelling*, W. J. Hardcastle and A. Marchal, Eds. Dordrecht: Kluwer Academic Publishers, 1990, pp. 187–209.

[2] K. N. Stevens, *Acoustic Phonetics*. Cambridge: MIT Press, 1998.

[3] A. Jongman, R. Wayland, and S. Wong, "Acoustic Characteristics of English Fricatives," *J. Acoust. Soc. Am.*, vol. 108, no. 3, pp. 1252–1263, 2000.

[4] B. Munson, E. C. McDonald, N. L. Deboe, and A. R. White, "The acoustic and perceptual bases of judgments of women and men's sexual orientation from read speech," *J. Phon.*, vol. 34, no. 2, pp. 202–240, Apr. 2006.

[5] J. Stuart-Smith, "Empirical evidence for gendered speech production: /s/ in Glaswegian," in *Laboratory Phonology 9*, J. Cole and J.-I. Haulde, Eds. Berlin: Mouton, 2007, pp. 65–86.

[6] E. Levon and S. Holmes-Elliott, "East end boys and west end girls: /s/-fronting in Southeast England," *Univ. Pennsylvania Work. Pap. Linguist.*, vol. 19, no. 2, pp. 111–120, 2013.

[7] R. J. Podesva and J. Van Hofwegen, "/s/exuality in small-town California: Gender normativity and the acoustic realization of /s/," in *Language, Sexuality, and Power: Studies in Intersectional Sociolinguistics*, E. Levon and R. B. Mendes, Eds. New York: Oxford University Press, 2015, pp. 168–188.

[8] H. K. Vorperian, S. Wang, E. M. Schimek, R. B. Durtschi, R. D. Kent, L. R. Gentry, and M. K. Chung, "Developmental Sexual Dimorphism of the Oral and Pharyngeal Portions of the Vocal Tract: An Imaging Study," *J. Speech, Lang. Hear. Res.*, vol. 54, pp. 995–1010, 2011.

[9] R. A. Fox and S. L. Nissen, "Sex-related acoustic changes in voiceless English fricatives," *J. Speech, Lang. Hear. Res.*, vol. 48, no. 4, pp. 753–65, Aug. 2005.

[10] P. Flipsen, L. Shriberg, G. Weismer, H. Karlsson, and J. McSweeny, "Acoustic Characteristics of /s/ in Adolescents," *J. Speech, Lang. Hear. Res.*, vol. 42, pp. 663–677, 1999.

[11] C. Tait and M. Tabain, "Patterns of gender variation in the speech of primary school-aged children in Australian English: the case of /p t k/," in *Proceedings of the 16th Australiasian International Conference on Speech Science and Technology*, 2016, pp. 65–68.

[12] I. Mees, "Patterns of Sociophonetic Variation in the Speech of Cardiff Schoolchildren," in *English in Wales: Diversity, Conflict and Change*, N. Coupland and A. R. Thomas, Eds. Clevedon: Multilingual Matters Ltd, 1990, pp. 167–194.

[13] Australian Bureau of Statistics, "2016 Census QuickStats: Yarrawonga, Victoria," 2017. [Online]. Available: http://www.censusdata.abs.gov.au/census_services/getproduct/census/2016/quickstat/SSC22904?opendocument. [Accessed: 22-Feb-2018].

[14] W. Labov, "Transmission and Diffusion," *Language (Baltim).*, vol. 83, no. 2, pp. 344–387, 2007.

[15] T. Kisler, F. Schiel, and H. Sloetjes, "Signal Processing via web services: the use case WebMAUS," in *Proceedings of Digital Humanities 2012*, 2012, pp. 30–34.

[16] J. Harrington, *Phonetic Analysis of Speech Corpora*. Malden, MA: Blackwell, 2010.

[17] R Core Team, "A language and environment for statisical computing. R Foundation for Statistical Computing." Vienna, Austria, 2014.

[18] K. Forrest, G. Weismer, P. Milenkovic, and R. N. Dougall, "Statistical analysis of word-initial voiceless obstruents: preliminary data.," *J. Acoust. Soc. Am.*, vol. 84, no. 1, pp. 115–123, 1988.

[19] A. Kuznetsova, P. Bruun Brockhoff, R. Haubo Bojesen Christensen, P. B. Brockhoff, and R. H. B. Christensen, "lmerTest Package: Tests in Linear Mixed Effects Models," *J. Stat. Softw.*, vol. 82, no. 13, pp. 1–26, 2017.

[20] R. Lenth, "Least-Squares Means: The R Package lsmeans," *J. Stat. Softw.*, vol. 69, no. 1, pp. 1–33, 2016.

[21] K. Whitman, "Looking out for the 'Aussie Bloke': Gender, Class and Contextualizing a Hegemony of Working-Class Masculinities in Australia," Unpublished PhD Dissertation, University of Adelaide, 2013.

Dialogue Acts in the AusTalk Map Tasks

Dominique Estival, Valeria Peretokina

MARCS Institute, Western Sydney University

[d.estival/v.peretokina]@westernsydney.edu.au

Abstract

This paper reports on the analysis of 12 Map Task dialogues from the AusTalk corpus, which were annotated for Dialogue Acts. We describe the annotation process, including the modifications to the SWBD-DAMSL coding scheme, and present the results of an analysis of the number, complexity, and function of turns across conversation role (Information Giver or Follower), gender, and gender dyad. Significant differences are found depending on the gender and role of the speakers and on the gender dyad. Overall, female speakers produce more turns than male speakers, and female-female dyads produce more multiple-label and mixed function turns.

Index Terms: speech corpus, Australian English, dialogue acts

1. Introduction

One of the components of the AusTalk corpus [1, 2] is the Map Task, a data gathering game aimed at collecting naturalistic conversational speech [3]. Each Map Task participant is given a map of the same environment. The map given to the information giver (IG) also shows a route that winds between landmarks and the IG must describe this route to the information follower (IF) who draws it on the other map. Landmark discrepancies between the maps are included to encourage IF-IG negotiation. Thus, the Map Task allows us to study discourse phenomena in spontaneous speech.

The AusTalk Map Task was designed for Australian English by Jette Viethen, Robert Dale and Felicity Cox, with landmark names chosen to elicit specific phonetic combinations of interest to researchers studying that language variety. In the AusTalk corpus, each of the 853 participants participated in two Map Tasks, once as the IG and once as the IF. This paper reports on the preliminary results of a project aiming to identify the Dialogue Acts (DAs) present in AusTalk Map Tasks and their linguistic features, building on [4, 5], and to investigate whether any gender, age, and dialect differences are observable in the participants' choice of DAs. Specifically, here we discuss the combined effect of conversation role and gender on the number, complexity, and function of the participants' turns.

2. Data and data processing

2.1. Data

Speech from 60 AusTalk Map Task recordings (up to 20 minutes each) had already been transcribed in the AusTalk Annotation project [6]. The transcriptions, as well as the audio data, are available from the Alveo Virtual Lab [7]. Twelve (12) of these 60 transcribed Map Tasks were selected for the current analysis to counterbalance speaker gender and recording location, and were then fully annotated for DAs. Table 1 presents the details of the 12 Map Tasks selected. More detailed information about the speakers is retrievable from the AusTalk demographic information collected at the time of recording in the speaker questionnaire (www. alveo.edu.au).

Table 1. *Transcribed AusTalk Map Tasks selected for analysis.*

Dyad	Site	Information Giver	Information Follower
female-female	USYD	4_1145	3_306
	UQ	2_534	3_377
	ANU	4_232	4_644
female-male	USYD	3_794	3_1035
	UQ	4_1277	1_738
	ANU	1_1119	3_1274
male-female	USYD	3_627	3_705
	UQ	4_68	4_1151
	USYD	3_42	4_923
male-male	USYD	3_112	1_322
	UQ	3_81	3_644
	ANU	1_178	2_113

2.2. Data processing

The AusTalk speech data had been originally transcribed with Transcriber [8] and the transcription files were in .trs format. There were at least two audio files per Map Task, one for the main speaker, who was the IG (Channel 6), and one for the second speaker, who was the IF (Channel 1). The audio files for the two channels were converted into one audio file readable by Transcriber, which was then opened with corresponding transcription file(s) and saved in Transcriber. Some transcription errors were corrected at that stage. The resulting .trs files were converted to Praat textgrids [9]. Two tiers were added to the textgrid files: one Dialogue Act label tier for the IG and one for the IF, as shown in Figure 1.

3. Coding the AusTalk Map Task dialogues

In the context of dialogues such as the Map Tasks, the coding is considered in terms of DAs rather than speech acts (SAs). A DA refers to a move within a dialogue, so it can link back to a previous utterance and it assumes there is a conversation. By contrast, a SA refers to the intention of the speaker, and does not assume there is a conversation or dialogue or even the existence of a listener (e.g., praying or swearing). DAs include questions, answers, directives, statements, and replies. The taxonomy of DAs is not based on the form of the utterances, but on their interpretation in context. For instance, an utterance with the syntactic form of a question is not necessarily a query (e.g., it can be a rhetorical question) and a query can be realised as a declarative as well as by a question (e.g., *"you want me to go right?"* or *"do you want me to go right?"*).

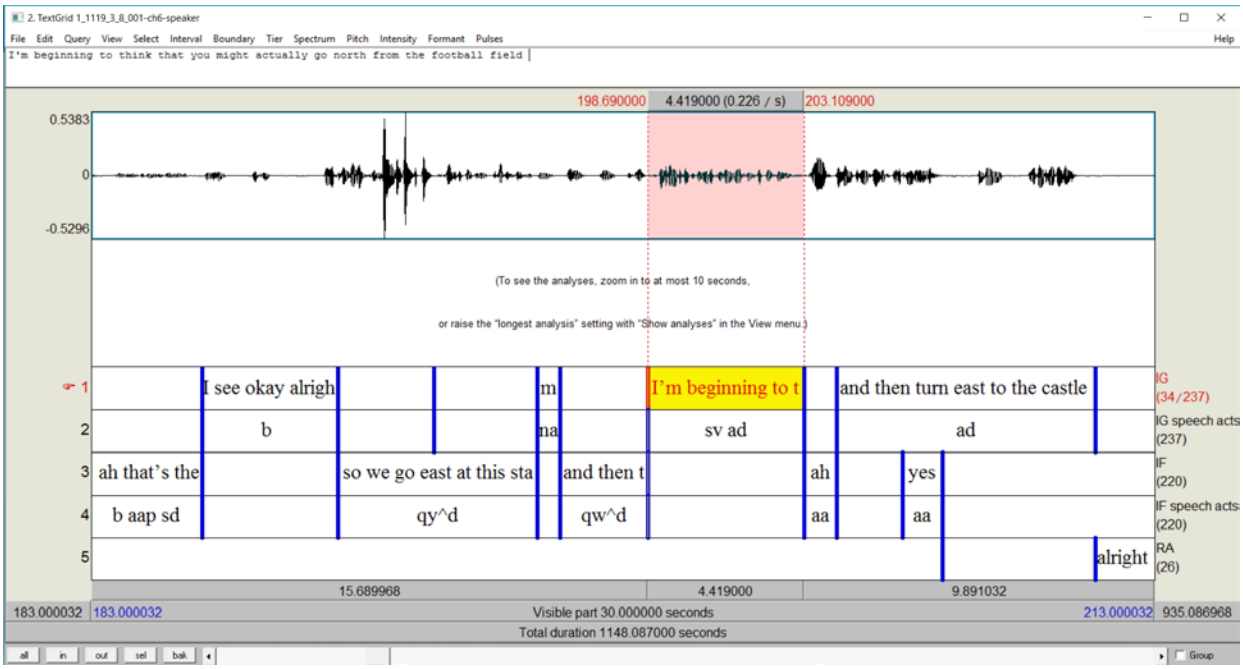


Figure 1. *Textgrid of AusTalk Map Task with orthographic transcription and Dialogue Acts labels for Information Giver (Tiers 1 and 2) and for Information Follower (Tiers 2 and 3). Tier 5 shows the transcription of the AusTalk Recording Assistant.*

3.1. Dialogue coding scheme

The coding scheme used in this project was adapted from the SWBD-DAMSL coding scheme [10], using insights from [5] and [4]. A useful aspect of DAMSL (Dialog Act Markup in Several Layers) is that it allows coding of dialogue moves at different levels at the same time, thus permitting the distinction between Task and Dialogue Management, as well as between backward- and forward-looking moves as follows:

- Task Management (e.g., request for clarification)
- Communication Management (e.g., clarification or apology)
- Backward-Communicative Function (e.g., clarification, confirmation or explanation)
- Forward-Communicative Function (e.g., follow-up or additional question).

In addition to the DA labels listed in [5], we adopted the following conventions for coding AusTalk Map Tasks DAs.

1) Questions

We specify whether a declarative question (*^d*) is a WH- or a Yes-No question, parallel to full interrogatives:

qy	yes-no-question	*Do you want me to turn left?*
qw	wh-question	*How far do I go?*
qy^d	declarative yes-no question	*We're going straight?*
qw^d	declarative wh-question	*You can find it where?*

2) Short responses (e.g., "*Yes.*", "*Right.*")

If the preceding DA was a question, the response is coded as *ny* for yes-answer, or *na* for affirmative non-yes answer. If the speaker is acknowledging a statement, the response is coded *aa* (accept/agreement).

3) Signal non-understanding (*br*) must be explicit (e.g., *"Excuse me?"),* rather than implied (e.g., *"mmm").*

4) *^e* (expansion) in *ny^e* (answer to yes-no question)

A simple expansion of a yes-answer (e.g., *"Yes, that's right*") remains *^e*. However, if a particular function can be identified, it is specified instead (e.g., *"Yep, and then you turn right"* is coded as *ny ad*: yes-answer + action directive).

5) *fo* (other forward function)

Anything used to propel the conversation forward which cannot be identified as any other label is coded as *fo*. An example of *fo* is *"So"* produced with long fricative and vowel.

6) *aa* (accept) versus *ny* (yes-answer)

IFs often produce *"yep", "yes", "Okay", "alright"* in response to action directives from IGs. If we identified the DA based on the lexical item, these responses would always be coded as *ny* or *na*. However, we view the DA function in combination with the previous utterance, that is, whether the interlocutor had asked a yes-no question or had been giving directions. Thus, in response to a question, *"yes"* is coded as *ny*. If it indicates that the IF is keeping up with the directions given to them (*"After the church you turn right - yep - and then go southerly"),* it is coded as *b* (acknowledge). If it indicates that the IF is agreeing with those directions (*"After the church you turn right - yep, got it"*), it is coded as *aa* (accept).

3.2. Inter-rater agreement

Three Map Tasks were coded independently by two research assistants. There was an average of 400 DAs per Map Task, 200 each for the IG and the IF, with an average of 36 DAs (16%) with different labels per Map Task. Manual examination showed most differences were systematic, with only 15 coding errors per Map Task (i.e., <4%). The systematic differences concerned the coding of:

- *aa* (accept) versus *b* (acknowledge) for utterances such as *"yes"* (see above);
- *nn^e* (no answer w/expansion) versus *nn* (no answer) for *"No, I don't"*;
- *oo* (open option) versus *ad* (action directives) for *"and we see a yellow church on the west side of the road"*.

4. Analysing Dialogue Acts

The 12 Map Tasks were analysed for differences in number of: turns; number of DAs per turn; relative frequency of backward and forward functions; relative frequency of complex and simple dialogue turns, according to the speakers' role in the conversation (IG vs. IF), gender (female vs. male), and the gender dyad completing the task (female-female, female-male, male-female, male-male). As Table 1 demonstrates, the chosen Map Tasks participant pairs consisted of three female-female dyads, three male-male dyads, and six mixed gender dyads. Of the mixed gender dyads, three had a female speaker as IG and a male speaker as IF, with the roles reversed (i.e., male IG and female IF) for the other three. For each participant, the total number of turns produced during the whole Map Task was calculated. Then, each turn was manually coded using the DA scheme described in Section 3.

Specific DAs and their variants were extracted from the textgrid by Praat scripts. Turns were grouped into two categories depending on the number of DAs identified in the turn. Turns coded with a single DA label comprised the single-label category, while turns containing multiple different DA labels were identified as multiple-label. In addition, the turns were categorised according to the function of their DAs: forward, backward, or mixed. Examples of different turn categories are presented in Table 2.

Table 2. *Turn categories.*

Category	Example	Labels assigned
Single-label	I want you to go past the orange castle should be on your right towards a yellow church with a tower but no steeple (3_42_4_10_IG)	*ad* (action directive)
Multiple-label	have we already been past this church yes okay (3_112_3_8_IF)	*qy* (yes-no question) *b* (acknowledge) *aa* (accept)
Forward	then we're going to head west till we get to New East Tunnel now that might be called something else (4_68_4_10_IG)	*ad* (action directive) *co* (offer)
Backward	okay just after the house (4_232_3_8_IF)	*b* (acknowledge) *^m* (repetition)
Mixed	oh okay so I turn right at the factory (1_178_4_10_IF)	*b* (acknowledge) *aa* (accept) *qy^d* (declarative yes-no question)

The aim of the present study was to determine whether the total number of turns, the frequency of occurrence of different types of turns (single- vs. multiple-label), and the dialogue function (forward vs. backward vs. mixed) varied depending on the participant's role in the Map Task, their gender, and the dyad type. There is no theoretical reason to assume other than the null hypothesis, except for the participant role (IG vs. IF) and the discourse function: we would expect that IGs would produce more forward function turns because they are giving directions to the IF. Conversely, IFs are expected to produce more backward function turns, either to acknowledge or question the information provided by the IG.

4.1. Results

One-sample chi-square tests were conducted to compare raw counts of: total turns, single-label turns, multiple-label turns, forward turns, backward turns, and mixed turns produced by the IG and IF participants. Similar tests compared the number of each type of turns produced by females and males, and by each dyad type. The alpha level for all tests was set at 0.05. Chi-square statistics are presented in Table 3.

Table 3. *One-sample chi-square statistics (cells containing significant results are highlighted in grey).*

	Role (df = 1)	Gender (df = 1)	Dyad (df = 3)
Total number of turns $N = 2{,}815$	$\chi^2 = 0.92$ $p = .336$	$\chi^2 = 74.84$ $p < .001$	$\chi^2 = 217.42$ $p < .001$
Complexity			
Single-label $N = 1{,}898$	$\chi^2 = 12.17$ $p < .001$	$\chi^2 = 35.07$ $p < .001$	$\chi^2 = 147.21$ $p < .001$
Multiple-label $N = 917$	$\chi^2 = 11.12$ $p = .001$	$\chi^2 = 44.06$ $p < .001$	$\chi^2 = 112.14$ $p < .001$
Function			
Forward $N = 957$	$\chi^2 = 260.19$ $p < .001$	$\chi^2 = 53.84$ $p < .001$	$\chi^2 = 113.17$ $p < .001$
Backward $N = 1{,}137$	$\chi^2 = 361.37$ $p < .001$	$\chi^2 = 8.97$ $p = .003$	$\chi^2 = 48.01$ $p < .001$
Mixed $N = 641$	$\chi^2 = 11.27$ $p = .001$	$\chi^2 = 14.08$ $p < .001$	$\chi^2 = 76.39$ $p < .001$

4.1.1. *Total number of turns*

According to the analyses, there was no statistical difference between the two roles (IG and IF) in the raw count of total number of turns used ($N_{IG} = 1{,}382$; $N_{IF} = 1{,}433$), confirming our hypothesis that total number of turns should not vary depending on role. However, the total number of turns was significantly different between genders as well as across the four dyad types. Female participants were found to produce more turns than male participants ($N_f = 1{,}637$; $N_m = 1{,}178$) and female-led dyads produced more turns than male-led dyads, with the highest number of total turns observed in female-female dyads ($N_{f\text{-}f} = 976$; $N_{f\text{-}m} = 803$; $N_{m\text{-}f} = 514$; $N_{m\text{-}m} = 522$).

4.1.2. *Turn complexity*

IGs were found to use significantly more multiple-label turns ($N_{IG} = 509$; $N_{IF} = 408$), forward turns ($N_{IG} = 728$; $N_{IF} = 229$), and mixed turns ($N_{IG} = 363$; $N_{IF} = 278$) than IFs. Conversely, IFs used single-label turns ($N_{IG} = 873$; $N_{IF} = 1{,}025$) and backward turns ($N_{IG} = 248$; $N_{IF} = 889$) more frequently than IGs. As female speakers produced more turns than male speakers overall, this difference remained statistically significant for both the number of single-label turns ($N_f = 1{,}078$; $N_m = 820$) and the number of multiple-label turns ($N_f = 559$; $N_m = 358$). Female-led dyads produced significantly more single-label turns ($N_{f\text{-}f} = 609$; $N_{f\text{-}m} = 604$; $N_{m\text{-}f} = 334$; $N_{m\text{-}m} = 351$) than male-led dyads. Interestingly, female-female dyads produced multiple-label turns more frequently than any other dyad type ($N_{f\text{-}f} = 367$; $N_{f\text{-}m} = 199$; $N_{m\text{-}f} = 180$; $N_{m\text{-}m} = 171$).

4.1.3. *Turn function*

As expected, IGs differed from IFs in their frequency of producing forward and backward function turns. The use of forward turns was more prevalent in the speech of IGs than IFs ($N_{IG} = 728$; $N_{IF} = 229$), with the opposite pattern found for the use of backward turns ($N_{IG} = 248$; $N_{IF} = 889$). It is interesting to note that the IG speakers also produced more mixed function turns than the IF speakers ($N_{IG} = 363$; $N_{IF} = 278$). Consistent with gender differences observed for the total number of turns

and the number of turns varying in complexity, females produced a higher number of turns of each function than males (forward turns: $N_f = 592$; $N_m = 358$; backward turns: $N_f = 619$; $N_m = 518$; mixed turns: $N_f = 368$; $N_m = 273$). Female-led dyads produced significantly more forward turns ($N_{f\text{-}f} = 342$; $N_{f\text{-}m} = 298$; $N_{m\text{-}f} = 155$; $N_{m\text{-}m} = 162$), and backward turns ($N_{f\text{-}f} = 346$; $N_{f\text{-}m} = 339$; $N_{m\text{-}f} = 221$; $N_{m\text{-}m} = 231$) than male-led dyads. Female-female dyads produced mixed function turns more frequently than any other dyad type ($N_{f\text{-}f} = 256$; $N_{f\text{-}m} = 130$; $N_{m\text{-}f} = 130$; $N_{m\text{-}m} = 125$).

4.1.4. Combined influence of Role and Gender

To disentangle the influence of conversation role from gender on the number and type of turns produced, a two-factor chi-square analysis was carried out. A significant relationship between role and gender was found for the total number of turns used in map tasks, $\chi^2(1, N = 2{,}815) = 31.41$, $p < .001$, for the number of single-label turns, $\chi^2(1, N = 1{,}898) = 48.84$, $p < .001$, and for the number of backward turns, $\chi^2(1, N = 1{,}137) = 47.88$, $p < .001$.

Females produced significantly more turns than males as IGs ($N_f = 877$; $N_m = 505$), but those gender differences were not as pronounced when males and females assumed the role of IFs ($N_f = 760$; $N_m = 673$). IG males produced significantly fewer single-label turns ($N = 302$) than IF males ($N = 518$), IG females ($N = 571$), and IF females ($N = 507$). Finally, IG females produced backward turns ($N = 183$) more frequently than IG males ($N = 65$), while the raw counts of this turn category was comparable (and, as expected, significantly higher) for IF males and females IFs ($N_f = 436$; $N_m = 453$).

5. Conclusions

This project contributes to the field of Dialogue Analysis by building upon the work of [5] and [4] and by further extending the SWBD-DAMSL Dialogue Coding Scheme [10] for Map Tasks. We refined the coding of questions by adding two labels for declarative questions, and specified the coding of short answers and expansions according to dialogue context.

The analysis of the distribution of DAs according to gender and Map Task role presented in Section 5 revealed interesting patterns which contribute to a better understanding of gender and conversation. No single independent variable (role, gender, and gender dyad) accounted for the communicative pattern variation observed in the data. Females did not conform to what we might expect from the conversational role: while producing more forward-function turns as IG than IF, they produced more backward-function turns than male IGs. Females also produced more turns and more complex turns than males, with this being even more pronounced in female-female dyads. Finally, IFs produced more single-label turns while IGs produced more multiple-label and mixed turns. Further investigation needs to be undertaken to explain these varied and complex data patterns.

Future work on this subset of AusTalk Map Tasks will examine the differences between speakers using total speaking time as a proxy for task success (all AusTalk Map Tasks were successful but they differed in length, with some pairs completing the task much more quickly than others) and turn complexity as a potential marker of difficulty in negotiating the map discrepancies. Another direction would be to analyse the differences between turns dedicated to task/conversation management and whether they also vary depending on the role, gender, and dyad.

In addition to providing preliminary observations on the influence of Map Task role and gender on the complexity and function of DAs in conversational speech, this project is a good example of what can be done with the AusTalk corpus, the audio-visual corpus of Australian English created by the Big ASC project [1]. This study built upon the Annotation Task of the Big ASC, which had produced transcriptions for selected components (both spontaneous and read speech) for a subset of the 853 AusTalk speakers; it also contributed to the data repository of the Alveo Virtual Lab, by adding specific annotations to the AusTalk corpus. Thus, an important outcome of the project consists of a revised set of transcriptions for 16 AusTalk Map Tasks, combining both IG and IF, and of DA annotations for 12 of those Map Tasks. As an early user of the new Alveo functionality allowing adding annotations to data held in Alveo, this project helped with the development and testing of the Alveo Annotation Contribution tool and provided feedback on the usability of the tool and on the interface.

6. Acknowledgements

The project was funded by a Transdisciplinary Grant from the Centre of Excellence for the Dynamics of Language (COEDL).

7. References

[1] Burnham, D., et al., *Building an audio-visual corpus of Australian English: large corpus collection with an economical portable and replicable Black Box*, in *Interspeech 2011*. 2011: Florence, Italy.

[2] Estival, D., et al., *AusTalk: an audio-visual corpus of Australian English*, in *9th Language Resources and Evaluation Conference (LREC 2014)*. 2014: Reykjavik, Iceland.

[3] Anderson, A.H., et al., *The HCRC Map Task Corpus.* Language and Speech, 1991. **34**(4): p. 351-366.

[4] Fletcher, J. and L. Stirling, *Prosody And Discourse In The Australian Map Task Corpus*, in *The Oxford Handbook of Corpus Phonology*. 2014, Oxford University Press.

[5] Fletcher, J., et al., *Intonational rises and dialog acts in the Australian English map task.* Language and speech, 2002. **45**(3): p. 229-253.

[6] Cassidy, S., D. Estival, and F. Cox, *Case Study: The AusTalk Corpus*, in *Handbook of Linguistic Annotation*, J. Pustejovski and N. Ide, Editors. 2017, Springer Verlag. p. 1287-1302.

[7] Cassidy, S. and D. Estival, *Supporting Accessibility and Reproducibility in Language Research in the Alveo Virtual Laboratory.* Computer Speech & Language, 2017. **45**: p. 375-391.

[8] Barras, C., et al., *Transcriber: development and use of a tool for assisting speech corpora production.* Speech Communication, 2000. **33**(1-2).

[9] Boersma, P. and D. Weenink. *Praat: doing phonetics by computer (Version 5.3.51)*. 2014; Available from: http://www.praat.org/.

[10] Jurafsky, D., L. Shriberg, and D. Biasca. *Switchboard SWBD-DAMSL Shallow-Discourse-Function Annotation, Coders Manual, Draft 13* 1997; Available from: https://web.stanford.edu/~jurafsky/ws97/manual.august1.html.

Varietal differences in categorisation of /ɪ e æ/ A case study of Irish and Australian English listeners in Melbourne

Chloé Diskin[1], Deborah Loakes[1,2] and Josh Clothier[1,2]

[1] School of Languages and Linguistics, The University of Melbourne
[2] ARC Centre of Excellence for the Dynamics of Language
{cdiskin, dloakes, joshuajc}@unimelb.edu.au

Abstract

This paper presents results of a vowel categorisation task of front lax vowels in /hVt/, /hVl/ and /mVl/ contexts, by 12 native Australian English speakers and 10 Irish migrants residing in Melbourne. Results show significant differences in how listeners categorise these vowels, in five out of six phonetic contexts. Vowels suggested to be undergoing merger in Victoria, specifically /el-æl/, are not perceived as merged, indicating this phenomenon may be stratified and/or more age-graded than previously reported. Results show clear differences between listeners sharing an L1 but speaking different dialects, even when these dialects are in direct contact due to migration.

Index Terms: second dialect acquisition (perception), vowel categorisation, Australian English, Irish English

1. Introduction

1.1. Background

Australia has long been known as a "classical immigration country", whereby the inflow of individuals taking up residence in Australia far outweighs the outflow, the numbers of individuals that leave Australia for elsewhere [1]. The larger cities of Australia are viewed as highly multicultural, 'superdiverse' [2], and multi-layered, with ongoing immigration resulting in diversification of urban areas, where new migrants settle amongst those who are more established. The number of migrants taking up residence in Australia is currently on the increase, particularly from Ireland, where there has been a long tradition of emigration, particularly to English-speaking countries. Between 2006 and 2014, there was a 39% increase in Irish-born people residing in Australia [3].

Despite increased contact due to migration, there exists a paucity of research into how speakers of the same L1 (English), but a different dialect, produce, process and perceive the sounds within these dialects. This study focuses on how native Irish English (IrE) speakers residing in Australia categorise /ɪ e æ/ in an experimental task. These vowels are of particular interest in the Australian English (AusE) context, as they are reported to have changed substantially over the years, raising "to a peak height" and then lowering again [4], to a point where [æ] is considered at the bottom of the vowel space [5]; see also [6], and can thus be challenging for native AusE listeners [7, 8]. Additionally, in Melbourne where the current study is carried out, a vowel merger is known to be in progress in the community, where /el/→/æl/. This pre-lateral context is thus also particularly variable in production and perception [7]. While previous research into perception of AusE by native listeners is limited, Mannell [8] has shown that the perception of these particular AusE vowels has certainly shifted over time, so that the perceptual boundary between /ɪ/-/e/ is higher than it once was, while for /e/-/æ/ it is lower. Mannell [8] attributes the shifted perceptual boundaries to accompanying diachronic production, reported in particular by Cox (e.g. [5]). The processing of vowel merger has also been shown to be highly variable, as well as being somewhat age-graded [7] and also geographically defined [9, 10].

1.2. Variability in second dialect processing

As noted by Sumner and Samuel [11] "listeners are confronted by a remarkably variable signal when they understand spoken language […] and a central issue in the perception of spoken language […] is the ways in which this variation is accommodated." While this is a complex issue for first language processing (see also [12]), even greater complexity is evident when listeners are faced with a second dialect, given the need to deal with multiple linguistic systems (cross-language perception is of course another related matter, but outside the scope of this study).

Second dialect acquisition, and its interaction with sociophonetics, gives us insight into how people cope with speech and communication in their new environment(s). Perception research in this area is relatively limited; but opens up a better understanding of "how sounds and words are learned, represented, processed and linked to social information" ([13]; see also references therein).

Evans & Iverson [14] discuss the issue of second dialect perception as one of potential accent normalisation, where "listeners may be able to fully adjust to vowel differences between accents, provided that they have had previous experience with similarly accented speech". They assessed how listeners of the same language (English), but of different British dialects, categorise vowels – to determine whether exemplar locations change due to the accent being listened to. Listeners from London heard two different accents and gave goodness ratings for vowel phonemes. Results were complex, but some crucial findings were that listeners were able to adjust their perception depending on the accent, and that in various ways participants' own backgrounds and production impacted responses. The finding regarding production correlates with a study by Allen & Miller [15], who found that "speech-specific experience" likely weighs heavily in guiding listeners to make choices about phoneme categorisation.

1.3. The 'Superdiversity' project

The vowel categorisation task in this study was completed by participants as part of a larger study examining second dialect and second language acquisition among two migrant groups (Irish and Chinese) in Melbourne, along with a 'control' group of native AusE speakers (see 2.2). The study included the recording of sociolinguistic interviews and wordlists, as well as the recording of ultrasound images of participants reading a wordlist. The vowel categorisation task was generally the last or second-to-last task to be completed by participants, and took approximately fifteen minutes to complete. While reaction times were recorded, they are not reported on here. While the study is in the early stages of analysis, we have previously found that in production (wordlists), the IrE female participants in this study have on average lower and more retracted front lax vowels than the AusE female participants [16], which may well have an impact on the way they in turn categorise vowels. It is important to note that there is to date no comparable research with IrE listeners who have never left Ireland.

1.4. Aims

The broad aims of this study are to compare how IrE listeners, living and working in Melbourne, categorise AusE vowel stimuli as compared to native listeners. As such, this study tests whether 'perceptual learning' or 'accent normalisation' has taken place, or whether IrE listeners are driven by exemplars from their own accents when categorising vowels. Using vowel continua, we interrogate the potential nuances within crossovers between vowel categories in perception, and what happens in different consonantal environments (i.e. how the listeners deal with coarticulation).

1.4.1. Research Questions

1. Are there differences in how Irish migrants and native AusE listeners categorise the lax front vowels /ɪ e æ/ (in 'control' condition /hVt/) produced by a native AusE speaker?

2. (a) How do Irish migrants and native AusE listeners respond to vowels in coarticulated contexts; prelateral /hVl/ and nasal onset prelateral /mVl/?

2 (b) Are there differences between the Irish migrant and AusE listeners in the categorisation of vowels known to be undergoing sound change (merger) in southern Victoria, specifically /el-æl/?

2. Method and analysis

2.1. Experimental task

The phonetic categorisation task is a forced-choice identification task, presented on an iPad using a specifically designed custom app. Individual words were played to listeners via headphones (Shure SRH840 Reference Studio Headphones) and items were also presented orthographically on 'buttons' on the screen. Of the two options presented, listeners made a choice by pressing which of the two items they had heard before moving on to the next item. To create the stimuli, seven-step continua were created using the Akustyk vowel synthesis module [17] in Praat [18]. The continua we use in this study all involve front lax vowels in various contexts, broadly /hVt/, /hVl/ and /NVl/, and include a mix of low and high frequency words, and include proper names. The experiment includes a number of back vowel stimuli, but we do not report on these here. The items used in this study are shown in Table 1 below.

Table 1. *The six continua analysed in this study. Predicted merger conditions (see [7, 9]) are bolded.*

Phonetic context	/i-e/	**/e-æ/**
/hVt/	*hit-het*	*het-hat*
/hVl/	*hill-hell*	***hell-hal***
/mVl/	*mill-Mel*	***Mel-Mal***

Listeners were timed for each item, and they could not replay the item, go backwards in the experiment, or change their mind once a decision was made. Each item was presented four times, with the orthographic representation shown twice on the left side of the screen, and twice on the right. Listeners were aware they were listening to AusE – the overall study was described to participants as being about the adoption of AusE.

2.2. Participants

In total, 12 Australian (7F, 5M) and 10 Irish (5F, 5M) took part in the study and completed the vowel categorisation task. The AusE speakers were all born and raised in Melbourne and had not spent any significant period of time (more than 1 year) outside Melbourne. The Irish migrants came from different towns and villages across the island of Ireland (North and South) and had migrated to Australia at various stages throughout the 2000s, with lengths of residence in Australia ranging from 1 to 14 years. Seven of the ten Irish migrants had only ever lived in Ireland and Australia; three had also lived elsewhere (UK: 2 years and 11 years; USA: 2 years). Year of birth among all participants ranged from 1976 to 1991, with the average age of the Australian group at 33, and for the Irish group at 35.

3. Results

3.1. Dialectal differences in /hVt/ perception.

Response curves for the control continua, fit using logistic functions in the *quickpsy* [19] package for *R* [20], are shown in Figure 1, with /ɪ-e/ on the left, and /e-æ/ on the right.

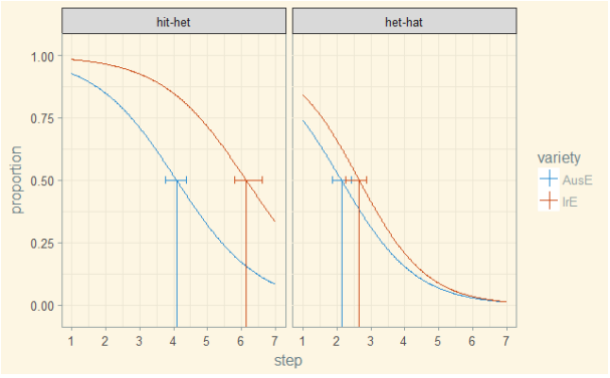


Figure 1. *Response curves for the control conditions /hit-het/ and /het-hat/*

Figure 1 shows that the AusE and IrE listeners have significantly different crossover points for /ɪ-e/, with the AusE listeners switching categories at Step 4 (the mid-point of the 7 stimuli) and IrE listeners crossing over much later at Step 6.

Exact crossovers are also shown in Table 2 further below. This suggests IrE listeners need an acoustically lower and more retracted vowel to classify a token as /ɪ/ compared to AusE listeners. Considering the endpoints of the /ɪ-e/ continua, 100% agreement was reached only for Step 1 of *hit* and only for the IrE listeners. 100% agreement was not reached for the AusE listeners for either end of the continua for this contrast. Furthermore, at Step 7 the majority of AusE listeners switch categories to /e/, but this is not true for the IrE listeners, who in many cases categorise Step 7 as /ɪ/.

For the /e-æ/ contrast, there is more agreement amongst the listener groups, in the sense that there is no significant difference in the crossover from /e/ to /æ/ (see also Table 2). Additionally, both listener groups have switched to hearing *hat* at Step 7. At Step 1, there is a lack of certainty for some listeners, so 100% agreement is never reached for *het*, and there is also an overall bias towards *hat.*

Table 2. *Modeled crossover points for /hVt/ control condition showing 95% upper and lower CI *indicates significant differences*

Contrast pair	Variety	50% cross-over	95% CI lower	95% CI upper
*hit-het**	AusE	4.09	3.73	4.37
	IrE	6.14	5.79	6.51
het-hat	AusE	2.13	1.68	2.46
	IrE	2.65	2.30	2.95

3.2. Coarticulatory effects on categorisation

Figure 2 shows that both IrE and AusE listeners have 100% agreement for *hill*. For the IrE group however, a *hill* response is sustained until around Step 5, while the AusE listeners have a gradual decline in *hill* responses until the category crossover at almost exactly Step 5. While the majority of AusE listeners hear *hell* at Step 7, there is still a large number of Irish listeners preferring *hill.*

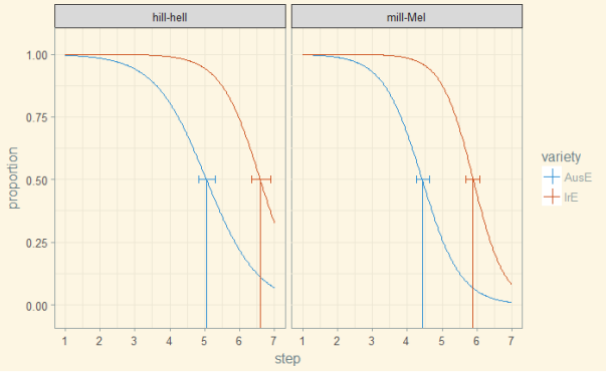


Figure 2. *Response curves for /hill-hell/ and /mill-Mel/*

For /*mill-Mel/,* responses are very similar to /*hill-hell*/, in that the IrE group sustain *mill* responses for longer than the AusE group, and again there are significant differences in the crossovers. As seen in both Figure 2 and Table 3, however, the crossovers are slightly earlier in this nasal onset condition.

Table 3. *Modeled crossover points for /ɪ-e/ continua in /hVl/ and /mVl/ showing 95% upper and lower CI *indicates significant differences*

Contrast pair	Variety	50% cross-over	95% CI lower	95% CI upper
*hill-hell**	AusE	5.05	4.85	5.29
	IrE	6.59	6.36	6.87
*mill-Mel**	AusE	4.43	4.24	4.60
	IrE	5.89	5.74	6.09

Continuing with the issue of coarticulation, and now including the 'merger' context, response curves for the /el-æl/ condition are shown in Figure 3. This shows significant differences between how the IrE and AusE listeners respond. For the */hell-Hal/* context the AusE listeners have a crossover at Step 4 (essentially the exact acoustic midpoint); whereas the IrE listeners have a very late crossover: close to Step 6 (see also Table 4). When a nasal is present, the crossovers are somewhat earlier; as seen in Table 4, it is half a step earlier for the AusE listeners and a full step earlier for the IrE listeners.

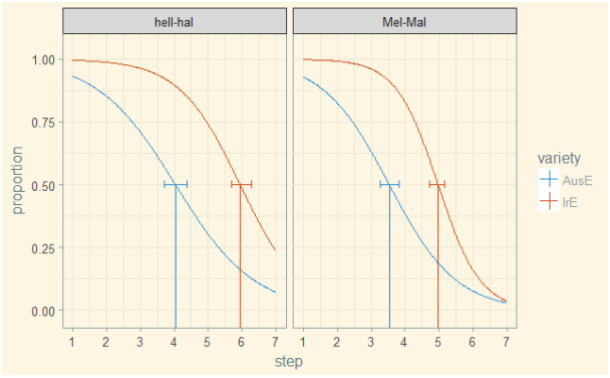


Figure 3. *Response curves for /hell-Hal/ and /Mel-Mal/*

The */hell-Hal/* context is dubbed the 'merger condition', but there is little evidence for a merger in perception among the AusE listeners, as well as (less surprisingly) for the IrE listeners. The endpoints of the continua show that the majority of AusE listeners hear *hell* at Step 1 and *Hal* at Step 7 – indicating very little ambiguity. Previous research showing a merger in perception shows listeners: (a) have a preference for *hell*; and (b) have a tendency for answering at random by point 7 (if [el] can be the same as [æl], then it follows that people cannot accurately categorise Step 7), [e.g. 7, 10]. While many of the IrE listeners are still answering *hell* by Step 7, these results tend to mirror the other continua analysed so far, with IrE listeners having a late crossover and lack of agreement at this stage, rather than a merger in perception.

Table 4. *Modeled crossover points for merger condition /el-æl/ showing 95% upper and lower CI *indicates significant differences*

Contrast pair	Variety	50% cross-over	95% CI lower	95% CI upper
*hell-Hal**	AusE	4.03	3.76	4.34
	IrE	5.94	5.70	6.26
*Mel-Mal**	AusE	3.53	3.21	3.78
	IrE	4.97	4.75	5.19

4. Discussion and conclusion

The results presented in this paper show differences in categorisation behaviour by IrE and AusE listeners when responding to AusE stimuli. The findings show that Irish participants tolerate a higher F1 and lower F2 (a more open and retracted vowel) before switching to the next category in each of the following pairs: *hill-hell, hell-Hal, mill-Mel, Mel-Mal* and *hit-het* – IrE listeners had significantly different crossovers from the AusE listeners in these cases. The *het-hat* contrast was an exception. Lack of certainty amongst the IrE and AusE listeners for /e/ in Step 1, and the early crossover for both listener groups, could be a word frequency effect (*het* is infrequent) or could be related to the fact that both listener groups simply require a phonetically higher vowel in this context. On the issue of the vowel crossovers for IrE and AusE listeners, we can say that while AusE vowels have lowered and retracted over time (e.g. [4]); in perception, the IrE cohort has even lower and more retracted vowels. For the female speakers at least, this corroborates with their production of close front vowels (i.e. [16]).

Addressing the effect of coarticulation on perception, the results of this study mirror expectations. For the AusE listeners, crossovers are always latest for /hVl/ contexts, and earliest for the control condition. The context with a nasal falls in between, indicating a "push and pull" between anticipatory and carryover coarticulation. For the IrE listeners this pattern is also observable, but only in the /e-æ/ condition. For /ɪ-e/, the nasal onset has the earliest crossover followed by the control condition, and finally the prelateral (crossovers are all quite close for the IrE listeners, however, all falling between 5.89 and 6.59). Despite varietal differences, for the cohort as a whole, the lateral lowers the vowel in perception ('downshifting' – category crossover is later) and the nasal onset raises the vowel in perception ('upshifting' – category crossover is earlier). Previous research on vowel merger in perception, using precisely the same research tool as this study, has shown varying degrees of ambiguity in categorisation of /el-æl/ in southern Victoria (see 1.1). However, in this study we have little evidence of merger by the native Melbourne listeners. While still to be independently verified, this may be highlighting an age-graded phenomenon, whereby younger listeners are now *not* merging in perception. This may have gone unreported previously as listener groups in earlier studies have been older than the participants here.

Finally, results show that the response curves work well for native AusE listeners, with the crossover points largely in the centre of Step 1 and Step 7. This is unsurprising as Step 4 is an acoustic midpoint synthesised from real speech stimuli by an AusE speaker. The *het-hat* context is an exception, with many of the listeners biased toward *hat,* as discussed. The vowel continua do not 'match' in the same way for IrE listeners (less agreement; very late crossovers). This finding gives support for the idea that listeners are very much guided by their own dialects in making vowel phoneme judgements [following 14,15]). The IrE listeners, despite listening experience (to varying degrees through living, studying and working with Australians), still 'hear' through their first dialect. Future research with these participants will test the effect of length of residence to see whether prolonged exposure to AusE results in changes in categorisation behaviour.

This study provides solid support for the idea that people's dominant dialect (in which they have received the most exposure, and which they evidently speak) influences their categorisation of vowels in a second dialect. The study also opens up the question of the amount of processing difficulty (and even possible misunderstandings) caused in a new dialect environment, despite the shared language. We will address this in future research, comparing processing times across the IrE and AusE cohorts, and with more fine-grained analyses relating to individual participants' own productions.

5. References

[1] S. Castles, *Ethnicity and Globalization: From Migrant Worker to Transnational Citizen*. London: Sage, 2000.

[2] S. Vertovec, *The Emergence of Super-Diversity in Britain*. Oxford: Centre on Migration, Policy and Society, 2006.

[3] Australian Government Department of Immigration and Border Protection, "Country Profile – Ireland", 2016. Available: https://www.homeaffairs.gov.au/about/reports-publications/research-statistics/statistics/live-in-australia/country-profiles/ireland

[4] F. Cox and S. Palethorpe, "Reversal of Short Front Vowel Raising in Australian English" in *Proceedings of Interspeech 2008*, Brisbane, 2008, pp. 342- 345.

[5] F. Cox, "The Acoustic Characteristics of /hVd/ Vowels in the Speech of Some Australian Teenagers", *AJL*, vol. 26, 2, pp. 147-179, 2006.

[6] J. Elvin, D. Williams and P. Escudero, "Dynamic acoustic properties of monophthongs and diphthongs in Western Sydney Australian English", *JASA*, vol. 140, 1, pp. 576-581, 2016.

[7] D. Loakes, J. Fletcher and J. Hajek, "Can you t[æ]ll I'm from M[æ]lbourne? Merger of the DRESS and TRAP vowels before /l/ as a regional accent marker in Australian English", *English World-Wide*, vol. 38, 1, pp. 29-49, 2017.

[8] R. Mannell, "Perceptual vowel space for Australian English lax vowels: 1988 and 2004" in *Proceedings of the 10th SST*. Macquarie University, NSW, 2004, pp. 221-226.

[9] F. Cox and S. Palethorpe, "The border effect: Vowel differences across the NSW – Victorian border" in *Proceedings of the 2003 Conference of the ALS,* Uni Newcastle, NSW, 2004, pp. 1–27.

[10] D. Loakes, J. Hajek, J. Clothier and J. Fletcher, "Identifying /el/-/æl/: a comparison between two regional Australian towns" in *Proceedings of the 15th SST*, Canterbury, 2014, pp.41-44.

[11] M. Sumner and A.G. Samuel, "Perception and representation of regular variation: the case of final /t/", *J Mem Lang*, vol. 52, pp. 322-328, 2005.

[12] A. Cutler, *Native listening: Language experience and the recognition of spoken words*. Cambridge, MA: MIT Press, 2012.

[13] J. Nycz, "Second dialect acquisition: A sociophonetic perspective", *Lang Ling Compass*, vol. 9, 11, pp. 469-482, 2015.

[14] B. G. Evans and P. Iverson, "Vowel normalization for accent", *JASA,* vol. 115, 1, pp. 352-361, 2004.

[15] J. S. Allen and J. L. Miller, "Contextual influences on the internal structure of phonetic categories", *Percept Psychophys*, vol. 3, 5, pp. 798-810, 2001.

[16] C. Diskin, D. Loakes and B. Volchok, "A sociolinguistic investigation of the adoption of Australian English by the Irish migrant community", in *Language Variation and Change-Australia*, University of Sydney, NSW, 2017.

[17] B. Plichta and D. Preston, "Akustyk for PRAAT" [Computer program], Version 1.7.2. East Lansing: Michigan State University, 2004.

[18] P. Boersma, and D. Weenink, "Praat: doing phonetics by computer [Computer program], 2018. Version 6.0.40. Available: http://www.praat.org/.

[19] D. Linares, J. López-Moliner, "quickpsy: An R Package to Fit Psychometric Functions for Multiple Groups", *The R Journal*, vol. 8, 1, pp. 122-131, 2016.

[20] R Core Team, *R: A language and environment for statistical computing (Version 3.4.3)*. Vienna, Austria: R Foundation for Statistical Computing, 2017.

Exploration algorithm for learning of sensorimotor tasks using sampling from a weighted Gaussian Mixture

Denis Shitov, Elena Pirogova*, Margaret Lech* and Tadeusz A. Wysocki^*

*School of Engineering, RMIT University, Melbourne Australia
^College of Electrical and Computer Engineering, University of Nebraska-Lincoln
denis.shitov@rmit.edu.au, elena.pirogova@rmit.edu.au
margaret.lech@rmit.edu.au, twysocki2@unl.edu

Abstract

This study presents a sampling efficient algorithm of a goal-directed exploration for learning complex non-linear sensorimotor mappings. The proposed generic approach uses sampling from weighted Gaussian Mixture Models (GMs) with both positive and negative weights that is shown to be an efficient way of searching in a non-linear space with multiple local minima. The simulations were performed by training the articulatory model to learn five distinct sounds of English vowels: [a], [e], [i], [o], [u]. The results demonstrated that after 400 iterations, the algorithm generated sounds with the competence values above 82% for all 5 vowels.

Index Terms: speech acquisition, speech modelling, articulatory synthesis, Gaussian mixtures

1. Introduction

Sensorimotor tasks belong to a family of tasks where both, sensory and motor coordination, are required. One of the most complex real-life examples is the ability of controlling the vocal tract for producing particular sounds. The significant challenges presented by this task are not only due to the high degree of freedom of the vocal tract and high dimensional sensory stimuli, but also due to the necessity to build the model of relationship between the motor and sensory space denoted here as X and Y respectively. According to the [1], strategies of learning this relationship, are based on defining the so-called "goals" in the sensory space Y that outperform another group of strategies where, at each time step the agent explores the motor space X. A common aspect of these different approaches is that in both cases, the agent has to sample points in a chosen space [1].

This study presents an efficient approach to perform this sampling with modifications that increase the speed of the learning process. It should be noted that this study doesn't propose a completely new view on the problem of learning sensorimotor tasks. It rather serves as a complementary approach that could be applied to the existing strategies in order to improve the efficiency of the searching process. In addition, despite the fact that this study is focused on the example of the articulatory control of the vocal tract for speech production, the proposed algorithm can be easily generalized and applied to various similar tasks. Speech acquisition is the task that all humans face during first years of their lives. It is believed that the phenomenon known as "babbling" is the crucial aspect of the speech learning process. During the period of babbling infants explore abilities of their vocal tract and develop strong connections between particular vocal tract configurations and sounds produced with these configurations [2]. It is not clear yet how infants achieve the goal of babbling. Developmental psychology studies reveal that the presence of ambient speech influences patterns of babbling [3]. In other words, one can assume that infants explore their vocal tract configurations not randomly, but based on the comparative listening experience between their babbling and the ambient language. Thus, many models of speech acquisition implement babbling algorithms to imitate the infant's learning process. Exploration of the articulatory space is a common method of acquiring articulatory configurations used in successful imitation of desired sounds [4], [5]. One of the most recent models of babbling has been presented in [6]. This model overcomes some of the limitations of the previous studies outlined in [7], [8] by applying the Principal Component Analysis (PCA) and the Linear Discriminant Analysis (LDA) to the highly dimensional acoustic space and generating a low-dimensional goal space. This paper describes a solution of similar accuracy, but significantly higher sampling efficiency (defined as a number of sampling operations per one run of the algorithm), which could be an important criteria in cases when sampling takes significant amount of time compared to the time spent on the whole algorithm to run.

2. Method

2.1. Generation of ambient speech samples

The focus of this study was to imitate five distinct English vowels: [a], [e], [i], [o], [u]. The reason for considering this set of sounds is that all of these are cardinal vowels when the tongue is in an extreme position, meaning that learning process of producing the cardinal vowels requires an extensive exploration in the whole motor space. Each vowel was represented by 100 audio-samples synthesized by VocalTractLab (VTL)[9]. In order to imitate natural speech diversity, Gaussian noise was added to the predefined shape of the VTL for the corresponding vowel. Therefore, shapes distribution, used for the synthesis of ambient vowel sound samples, were given as $X \sim \mathcal{N}(\mu, \Sigma)$, where μ denotes the predefined shape of vowel for JD2, and Σ is a diagonal covariance matrix given as $\Sigma = \sigma^2 I$, where σ^2 denotes the variance for all variables. JD2 is the configuration of the male vocal tract with specified anatomical parameters and predefined settings for distinct speech sounds of the German language. For the purpose of low-dimensional representation of the ambient speech, 13 Mel-frequency cepstral coefficients (MFCC) were estimated for each ambient speech sample as in [6]. Despite the relatively low dimensionality of this representation it has been successfully used in speech processing for many years. Then for each vowel, an average MFCC vector y^* was

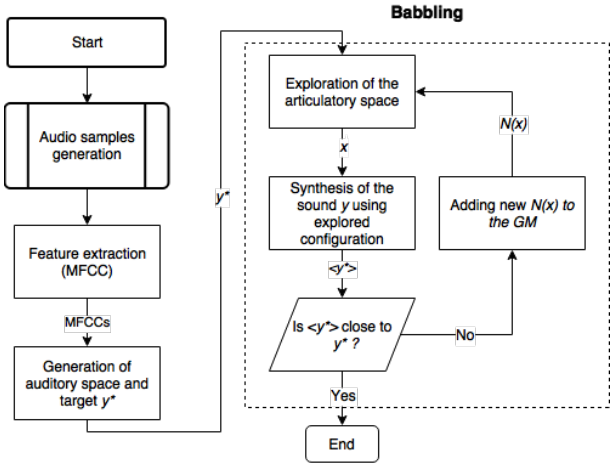


Figure 1: *Flowchart of the "babbling" procedure.*

calculated over all 13-dimensional feature vectors and used as a target for the model. This process can be described as:

$$y^* = \frac{\sum_{y \in Y} y}{|Y|}, \tag{1}$$

where y^* - target vector, Y - sample set representing a particular vowel and $|Y|$- size of the sample set Y.

2.2. Babbling procedure

Given the target y^*, the goal of the model was to learn how to imitate, or reproduce this target using a feed forward articulatory synthesizer. In other words, the aim was to find the vocal configuration x^* that produced the target sound y^*: $f(x^*) = y^*$, where $f(x)$ is the feed forward articulatory synthesizer's function. As mention before, the VTL synthesizer was used in this study. A number of previous studies proposed different approaches to the task of x^* estimation. This study is focused on one of the naturally-plausible methods assuming that babbling can be described as a process of iterative exploration of the articulatory space and development of connections between the articulatory and the auditory spaces. The convergence of this process is controlled by the auditory feedback, where the infant compares his own sounds with the memory of ambient sound samples. In technical terms, the goal of babbling can be formulated as a process of finding an inverse function $g(y^*)$ of the feed-forward synthesizer: $g(y) = x$. A general scheme for the described procedure is shown in Fig. 1, where $N(x)$ is a Gaussian distribution. At each iteration, the babbling procedure could be split into two stages, exploration and adaptation. The aims of the proposed approach are to: (i) Encourage the model to explore regions, where previous explorations have shown results close to the target y^*; and (ii) Inhibit the exploration of regions with negative exploration results.

In order to achieve highlighted aims, we suggest to introduce a function $h(x)$ strongly correlated with the probability density function $p(x)$ of the exploration configuration x: $p(x) \approx h(x)$. Given the high level of correlation between $p(x)$ and $h(x)$, $h(x)$ could be used to sample vectors x from the articulatory space during the exploration stage. These vectors could be then passed to the VTL to synthesize speech sounds. Both, the form of the sampling function $h(x)$ and the sampling algorithm, are considered to be the key contributors of this study. These are further explained in the following sections.

2.3. Proposed exploration algorithm

We suggest to represent $h(x)$ as a Gaussian Mixture (GM) where, each Gaussian component $\mathcal{N}(x|\mu_i, \Sigma_i)$ is added to the GM as a result of a single babbling iteration:

$$h(x) = \sum_{i=1}^{N} w_i \mathcal{N}(x|\mu_i, \Sigma_i) + w_0, \tag{2}$$

where N is the number of iterations, w_i is the weight of the i-th component that could be both positive and negative, μ_i, Σ_i are the mean and the covariance matrix of the i-th component respectively, and w_0 denotes the initial state of the system when all configurations x are equally likely to be selected for the exploration. For the sake of simplicity, an isotropic multivariate Gaussians were considered, meaning that the covariance matrix was a matrix with identical values on the main diagonal (i.e. $\Sigma = \sigma^2 I$). In regards to the Gaussian components weights w_i, it was important to ensure that the model allows for both, positive and negative weight values. This in turn, allows the model to either decrease or increase a probability of exploration in some regions. This property is likely to facilitate more efficient exploration of the articulatory space that can lead to faster search for the best values of x^*.

Despite the fact that having negative weight values was considered to be advantageous in terms of the search efficiency, it was not obvious how to sample the vectors x to maintain the probability distribution given by the function $h(x)$. Hence, an advanced sampling algorithm has been considered. The following modified version of the Rejection method described in [10] was proposed. At the k-th iteration the PDF function $h(x)$ is given as in (2), where $N = k - 1$ denoted the number of Gaussian components. The proposed algorithm proceeds as follows:

Step 1. Filter all components with non-negative weights: $w \in W_{pos}, \forall w \geq 0$.

Step 2. Calculate cumulative distribution function (CDF) denoted by $F(w)$ for W_{pos}

Step 3. Sample a random value rnd_0 from the uniform distribution in the range $rnd_0 \in [0, max(F(w))]$.

Step 4. Find the index i that: $F(w_{i-1}) < rnd < F(w_{i+1})$.

Step 5. Sample a random configuration x_{rnd} from the i-th component of the $h(x)$.

Step 6. Calculate PDF of x_{rnd} given only non-negative components of the GM: $h_{pos}(x_{rnd})$.

Step 7. Calculate PDF of x_{rnd} given only negative components of the GM: $h_{neg}(x_{rnd})$.

Step 8. Sample a random value rnd_1 from the uniform distribution in the range $rnd_1 \in [0, h_{pos}(x_{rnd})]$.

Step 9. Accept x_{rnd} if $rnd_1 > |h_{neg}(x_{rnd})|$, otherwise go to Step 3.

Fig. 2 shows an example of sampling using the proposed algorithm from a randomly generated distribution with six Gaussian components, where each dotted trace represents a corresponding gaussian component with additional straight line reflecting a prior uniform component. The solid contour illustrates the sampling function $h(x)$ and the bars show the histogram of the x values sampled by the proposed algorithm. According to the aim of the proposed method, the PDF of the sampled distribution follows the contour of $h(x)$ for positive values of $h(x)$.

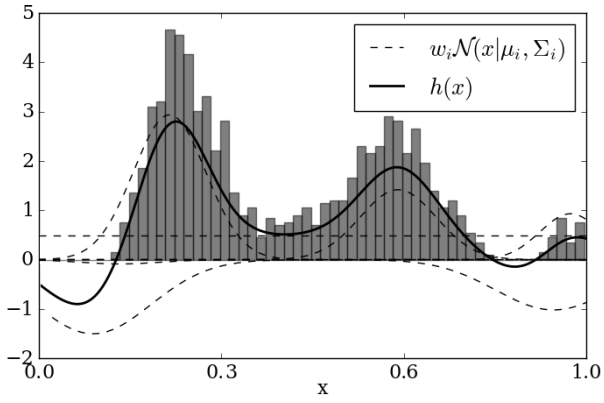


Figure 2: *An example of sampling from a randomly generated GM with $k = 6$ components.*

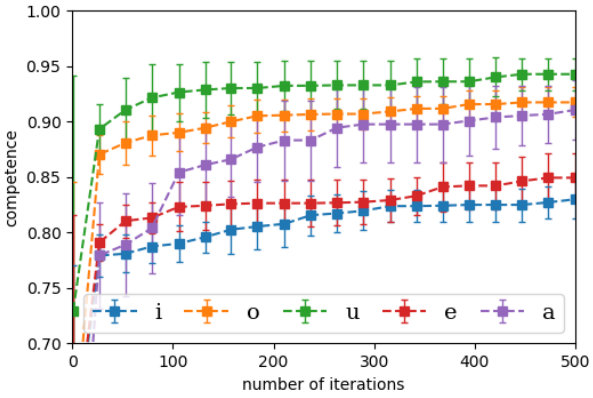


Figure 3: *Competence values (averaged over 10 trials) resulting from learning 5 vowels .*

2.4. Adaptation stage of babbling

Once the articulatory configuration x was sampled by the algorithm, it was used to produce a sound by the feed-forward model $f(x)$, and calculating the MFCC feature-vector y characterizing the sound acoustics. The pairs of vectors (x, y) were then passed to the adaptation step. The goal of the adaptation process was to evaluate how closely the estimated vectors y and the target vectors y^* were positioned within the auditory space. Euclidean distance measure d was used to determine the distance between parametric representations of these two sounds:

$$d(y, y^*) = \|y - y^*\|_2 = \sqrt{\sum_{i=1}^{n} (y_i - y_i^*)^2} \quad (3)$$

Given the distance d, a new Gaussian component was added to the GM with the weight $w(d)$ estimated based on the value of d:

$$w(d) = \frac{2}{(1 + e^{2*(d-C)})} - 1 \quad (4)$$

The formula given in (6) represents an inverted and shifted logistic (sigmoidal) function.The idea of using this formula was to encourage the model to explore more thoroughly regions around x that were within radius distance d smaller then C. This means that, areas of the parametric space that were less likely to provide optimal solutions were explored less thoroughly, thus reducing the overall number of computations. The mean value μ of a new component was kept constant for all components. At each iteration the "best match" $\langle y^* \rangle$ estimation of y^* was updated only if a new acoustic vector y was closer to the target vector y^* than the current best match i.e.:

$$\langle y^* \rangle = y_i, d(y_i, y^*) < d(\langle y^* \rangle, y^*) \quad (5)$$

3. Experiment

The exploration and adaptation stages described in the previous sections, have formed a complete single babbling iteration. It is expected that after a sufficient number of iterations, the algorithm should be able to find an articulatory configuration $\langle y^* \rangle$ providing the best match for a given target configuration y^*. To determine how close these two configurations are correlated, the competence parameter was calculated as follows:

$$comp(\langle y^* \rangle, y^*) = e^{-\|\langle y^* \rangle - y^*\|} \quad (6)$$

For each of the 5 vowels, the learning procedure was executed 10 times. Each trial consisted of 500 babbling iterations. The learning outcomes averaged over 10 trials and are shown in Figure 3. In this experiment, no stop condition within the training process was implemented. In future studies, an adaptive algorithm for stopping the training will be considered. It should be mentioned that this experiment is mainly focused on observing the dynamics of the training rather then estimating the final accuracy of the model. The contours show how the model competence is improved with an increasing number of iterations for each vowel. It can be seen that only after 100 iterations the competence for four out of five vowels become greater than 0.80, and after 400 iterations - reached a plateau with the competence above 0.82 for all five vowels. The competence results, reported in a related recent study [6] (we considered this algorithm as a conventional way of exploration) using a different approach, did not fall below 0.9 after 400 iterations. It should be mentioned that in the former, each iteration requires sampling of a batch with 10 targets, whereas our algorithm performs only one sampling per iteration. Competence scores in [6] were calculated over a 2-dimensional goal space, while in our study it was determined over a 13-dimensional space of the MFCC parameters, which makes direct comparison between scores irrelevant. Consistent with [6], the learning outcomes shown in our study varied slightly across different vowels with [u] achieving the highest results followed by [o], [a], [e], and [i] indicating that different vowels may need different number of exploration. Thus, [a] was gradually improved during the whole training process indicating that this specific sound can be produced in many different ways, in other words demonstrating many-to-one projection from motor to auditory space.

4. Conclusion

In this paper, a new modelling approach of the infant's babbling stage of speech acquisition was investigated. The presented learning process was based on random sampling of the articulatory configurations from the weighted GMs. While the positive components of the GMs encouraged the model to explore regions related to the target configurations, the negative components of the GMs discouraged exploration of regions loosely related to the target configurations. The experiments demonstrated that after 500 learning iterations, the proposed model

generated sounds of 5 vowels, namely [a], [e], [i], [o], [u] with the competence score exceeding 82%. Future study will investigate integration of the babbling procedure into modeling of more complex components of speech.

5. References

[1] C. Moulin-Frier and P.-Y. Oudeyer, "Exploration strategies in developmental robotics: A unified probabilistic framework," pp. 1–6, 2013.

[2] P. K. Kuhl, "Early language acquisition: cracking the speech code," *Nature reviews neuroscience*, vol. 5, pp. 831–843, 2004.

[3] A. C. B. Mampe, A. D. Friederici and K. Wermke, "Newborns' cry melody is shaped by their native language," *Current biology*, vol. 19, pp. 1994–1997, 2009.

[4] G. Westerman and E. R. Miranda, "Modeling the development of mirror neurons for auditory-motor integration," *Journal of New Music Research*, vol. 31, no. 4, pp. 367–375, 2002.

[5] J. J. S. M. Rolf and M. Gienger, "Goal babbling permits direct learning of inverse kinematics," *IEEE Transactions on Autonomous Mental Development*, vol. 2, no. 3, pp. 216–229, 2010.

[6] A. K. Philippsen, R. F. Reinhart, and B. Wrede, "Goal babbling of acoustic-articulatory models with adaptive exploration noise," in *2016 Joint IEEE International Conference on Development and Learning and Epigenetic Robotics (ICDL-EpiRob)*, Sept 2016, pp. 72–78.

[7] H. Liu and Y. Xu, "Learning model-based f0 production through goal-directed babbling," *International Symposium on Chinese Spoken Language Processing(ISCSLP)*, pp. 284–288, 2014.

[8] C. Moulin-Frier and P.-Y. Oudeyer, "Curiosity-driven phonetic laerning," *IEEE Intern. Conf. on Development and Learning*, 2012.

[9] P. Birkholz, "Modeling consonant-vowel coarticulation for articulatory speech synthesis," *PLOS ONE*, vol. 8, no. 4, pp. 1–17, 04 2013. [Online]. Available: https://doi.org/10.1371/journal.pone.0060603

[10] L. Devroye, *Non-Uniform Random Variate Generation*. Springer, 2013.

A comparative study on acoustic and modulation domain speech enhancement algorithms for improving noise robustness in speech recognition

Belinda Schwerin and Stephen So

School of Engineering and Built Environment,
Gold Coast Campus, Griffith University, QLD, 4222.
{b.schwerin, s.so}@griffith.edu.au

Abstract

This paper investigates whether modulation domain speech enhancement methods are better than corresponding acoustic domain methods when used as a preprocessor to automatic speech recognition. It is well known that linguistic information of speech is contained not only in the short-time magnitude spectrum but also in its temporal evolution. In addition, this study investigates whether popular metrics used in speech enhancement (such as PESQ, segmental SNR, STOI) are indicative of ASR performance. ASR experiments on the TIMIT speech corpus corrupted by various noises were performed to compare recent modulation domain methods with their acoustic domain variants.

Index Terms: modulation domain, robust speech recognition, speech enhancement

1. Introduction

It is commonly known that automatic speech recognition (or ASR) systems trained on clean speech will perform poorly when applied to speech that has been corrupted by environmental noise, in so-called mismatched conditions. A number of approaches have been investigated to mitigate the degrading effects of noise in ASR systems and these have been reported widely in the speech literature [1]. One popular approach is to apply a speech enhancement-based preprocessor on the noisy speech before it is passed to the ASR system. The premise for this approach is to use a speech enhancement algorithm to reduce the level of noise in order to improve the quality and intelligibility of the speech, which should assist in improving the ASR performance.

A large number of speech enhancement algorithms have been reported in the literature. We can classify all of these algorithms into two categories, based on the domain they process in. Typically, the short-time Fourier transform (or STFT) of the speech signal is processed, where the speech is windowed into short frames and a discrete Fourier transform is computed for each frame. In *acoustic frequency domain methods*, such as spectral subtraction [2], Wiener filtering [3] and MMSE-STSA (minimum mean squared error-short time spectral amplitude) [4, 5], the estimation is performed on the magnitude or power spectrum across all acoustic frequencies within each frame. A notable characteristic of acoustic frequency enhancement methods is that they generally process each short time frame independently without exploiting inter-frame dependencies that model the temporal dynamics of speech. For *modulation domain methods*, the estimation is performed on the modulation frequencies within the time trajectory of the magnitude [6] or the real/imaginary parts [7] of the STFT at each acoustic frequency; they are able to enhance the temporal dynamics of the power spectrum of the speech. Several modulation domain speech enhancement algorithms, such as modulation domain spectral subtraction [8], MMSE modulation magnitude estimation [9], and the modulation domain Kalman filter [10] have been reported to outperform their acoustic frequency domain analogues in terms of the quality of the enhanced speech.

In this paper, we investigate whether the advantage offered by recent modulation domain speech enhancement algorithms that are tuned for human listening, also results in improved performance in hidden Markov model (HMM)-based speech recognition. The feature vectors used in typical ASR systems comprise a parametric representation of the power spectrum, such as Mel-frequency cepstral coefficients (MFCCs) [11], as well as their first and second derivatives (also known as delta and delta-delta coefficients) [12], in order to exploit the temporal movements of the vocal tract. These derivatives have been shown in [13] to be equivalent to applying bandpass filters on the time sequences of spectral parameters (or TSSPs). Therefore, modulation domain enhancement methods have the potential to provide a better set of feature vectors for the ASR system.

Previous studies on modulation domain-based preprocessing in ASR, such as RASTA IIR filtering [6] and FIR-Slepian bandpass filtering [13], have demonstrated improvements in ASR accuracy when using basic filtering techniques in the modulation frequency domain. Therefore, this study examines the ASR performance when the noisy speech is preprocessed by recent and more sophisticated modulation domain speech enhancement algorithms. ASR experiments were performed on the TIMIT speech corpus [14] using the HMM Toolkit (or HTK) [15], that compare the performance across several acoustic and modulation domain enhancement methods for the white, F16 and babble noises. Phoneme correctness results from these experiments are presented along with speech quality (PESQ and segmental SNR) and intelligibility metric (STOI).

2. Method

In this paper we aim to evaluate the effect of applying modulation domain and RI (real and imaginary)-modulation domain based speech enhancement methods in the preprocessing stage, on the recognition rates of ASR. For this purpose, noisy stimuli were processed using various modulation domain, RI-modulation domain, and (for comparison) acoustic domain enhancement algorithms, then ASR experiments were conducted on these preprocessed speech stimuli. Details of these experiments are described below.

2.1. Speech corpus

The TIMIT speech corpus [14] was used for the ASR experiments. This corpus consists of 6300 utterances recorded from 630 different male and female speakers. The dataset is sampled at 16 kHz, and divided into training and testing sets. The training set consists of 3696 clean utterances from 462 speakers. The core test set, consists of 192 utterances from 24 speakers. Clean stimuli of the test set were corrupted with various noise types at input SNRs ranging from 0 dB to 20 dB. Noise types investigated include white (AGWN), babble, F16 and factory noises, and were generated with use of noise samples from the NOISEX-92 noise corpus [16]. Noisy test stimuli were processed by each speech enhancement method before their use in the ASR experiments.

2.2. Speech enhancement algorithms

The speech enhancement algorithms that were investigated include spectral subtraction, minimum mean-square error amplitude estimation, and Kalman filtering. Methods were implemented in the modulation domain, the RI-modulation domain, and for comparison in the acoustic domain. A total of 8 different enhancement methods were considered. Table 1 summarises each speech enhancement method evaluated, along with key parameters used in their implementation. Parameters applied for each are consistent with those given by the cited reference work and/or implementation.

2.3. ASR experiments

Automatic speech recognition (ASR) experiments made use of the TIMIT speech corpus (see section 2.1). The ASR model was generated using clean stimuli from the training set only. To prevent the biasing of the results, we removed the *sa** utterances from both the training and testing sets, as was done in [1]. For testing, noise corrupted stimuli from the core test set were first processed by each enhancement method described in section 2.2. Recognition tests were then conducted for each noise type, input SNR, and enhancement method type.

ASR experiments were conducted using an HTK-based triphone recogniser. Three states per HMM and 8 Gaussian mixtures per state were used. Consistent with [21], the set of 48 phonemes was reduced to 39 for testing. A frame size of 25 ms, and frame shift of 10 ms was used. MFCC features, energy coefficients, and first and second order derivatives were used, to give 39 coefficients in total. Cepstral mean subtraction was also applied. The bigram language model was used. Recognition used the Viterbi decoder, with no pruning factor, likelihood scaling factor of 8 and a penalty of 0. Recognition rates of phonemes were determined for each noise, input SNR, and treatment type in terms of correctness (Corr %).

3. Results and discussion

Results for ASR experiments for each noise type are shown in Table 2. Objective evaluation of the enhanced utterances used in each experiment are also shown in terms of mean PESQ score. Table 3 shows mean segmental SNR and STOI for each treatment type.

For higher SNRs, LOGAcMME was shown to be very effective in improving speech recognition rates. However, for lower SNRs, the results were less clear. When dealing with white noise, AcKal, a method originally designed for compensating AWGN corrupted stimuli, resulted in the highest recognition performance at all SNRs. RISSUB and ModSSUB outperformed AcSSUB, and LOGRIMME and LOGMME outperformed LOGAcMME at lower SNRs. For utterances corrupted with F16 noise at low SNRs, we similarly found that processing with LOGRIMME resulted in better recognition rates than when LOGAcMME was used. When considering babble noise, LOGAcMME was in general found to be the highest performing method. However, for the spectral subtraction-based methods, MdSSUB and RISSUB were found to outperform AcSSUB. This trend was also noticed in the Kalman filters, where MdKal outperformed AcKal. The recognition accuracies for babble noise indicated some degree of consistency with PESQ scores shown in Table 2. However, in general for the other noise types, it was found that high PESQ was not indicative of good recognition accuracy. This was expected since PESQ was originally developed for measuring perceptual speech quality in speech coding applications for human listeners.

The intelligibility metric, short-time objective intelligibility measure (or STOI) [22], on the other hand, consistently gave preference to MdSSUB. On the other hand, segmental SNR gave preference to LOGMME for babble noise, and either LOGMME or LOGRIMME for white and F16 noise types. These results highlight the difference between various metrics and ASR recognition rates, and the difference between methods yielding improved human listener preference and those yielding better ASR rates.

Considering the results reported for LOGMME and LO-

Table 1: *Enhancement methods evaluated for preprocessing stage of ASR. Important parameters used for each method are also given. These include Acoustic frame duration (AFD), Acoustic frame shift (AFS), Modulation domain frame durati on (MFD), Modulation frame shift (MFS), Smoothing factor α, number of Linear Prediction Coefficients (LPCs) used to model speech p , and the number of LPCs used to model noise q.*

Method	Implementation details
AcSSUB	Acoustic domain spectral subtraction [2] AFD = 20 ms, AFS = 10 ms, Power spectral subtraction.[17]
MdSSUB	Modulation domain spectral subtraction [8] AFD = 32 ms, AFS = 8 ms, MFD = 256 ms, MFS = 32 ms, Power spectral subtraction
RISSUb	RI-modulation spectral subtraction [18] AFD = 25 ms, AFS = 2.5 ms, MFD = 120 ms, MFS = 15 ms, Magnitude spectral subtraction
LOGAcMME	Acoustic MMSE Log-amplitude estimation [5] AFD = 20 ms, AFS = 10 ms, $\alpha = 0.98$ [17]
LOGMME	Modulation MMSE Log-amplitude estimation [9] AFD = 32 ms, AFS = 1 ms, MFD = 32 ms, MFS = 2 ms, $\alpha = 0.996$
LOGRIMME	RI-modulation MMSE Log-amplitude estimation [19] AFD = 32 ms, AFS = 1 ms, MFD = 32 ms, MFS = 2 ms, $\alpha = 0.996$
AcKal [20]	Acoustic domain Kalman filtering AFD = 50 ms, AFS = 6.25 ms, $p = 20$, $q = 10$.
MdKal [10]	Modulation domain Kalman filtering AFD = 32 ms, AFS = 4 ms, MFD = 40 ms, MFS = 40 ms, $p = 4$, $q = 8$.

Table 2: *TIMIT experimental results: mean PESQ scores and phoneme correctness (%) scores for babble, F16, factory, and white noises (clean corr = 75.82%). Highest scores are in bold.*

Algorithm	Mean PESQ					Corr (%)				
SNR (dB)	0	5	10	15	20	0	5	10	15	20
Noisy (babble)	1.75	2.10	2.45	2.79	3.13	24.64	34.81	46.16	57.66	66.24
AcSSUB	1.74	2.21	2.64	3.06	**3.44**	26.01	35.76	47.18	54.21	62.44
MdSSUB	1.99	**2.38**	2.73	3.04	3.33	33.58	43.05	53.65	62.79	69.08
RISSUB	1.88	2.31	2.70	3.05	3.37	31.21	40.48	51.52	61.43	67.43
LOGAcMME	**2.01**	**2.38**	**2.74**	3.08	3.39	**34.46**	**45.13**	**55.04**	**63.39**	**69.32**
LOGMME	1.87	2.31	2.72	**3.11**	**3.44**	32.48	39.80	49.33	58.15	65.29
LOGRIMME	1.90	2.33	2.73	**3.11**	**3.44**	32.03	40.10	49.67	58.62	65.70
AcKal	1.92	2.22	2.51	2.83	3.13	31.46	37.49	43.95	50.31	55.29
MdKal	1.88	2.26	2.62	2.95	3.25	31.58	40.23	49.76	58.29	64.06
Noisy (F16)	1.64	2.01	2.37	2.73	3.08	16.48	27.65	39.53	53.27	63.11
AcSSUB	1.91	2.38	2.84	**3.28**	**3.66**	28.52	40.02	50.75	58.76	64.98
MdSSUB	**2.32**	**2.63**	2.91	3.17	3.41	38.47	47.86	57.43	63.94	68.34
RISSUB	2.15	2.52	2.85	3.15	3.43	38.74	46.41	56.36	61.75	67.08
LOGAcMME	2.24	2.61	**2.94**	3.23	3.51	35.56	48.38	**59.30**	**66.10**	**69.27**
LOGMME	2.05	2.45	2.87	3.24	3.56	38.21	48.47	57.59	62.35	65.42
LOGRIMME	2.14	2.54	2.92	3.26	3.56	**40.68**	**50.70**	58.66	62.51	65.77
AcKal	2.13	2.51	2.88	3.20	3.47	39.34	49.00	56.13	62.49	65.86
MdKal	2.10	2.45	2.75	3.01	3.27	35.54	45.57	53.29	58.37	62.93
Noisy (white)	1.37	1.71	2.09	2.46	2.83	11.39	20.57	30.58	43.24	54.65
AcSSUB	1.69	2.19	2.66	3.11	**3.52**	24.51	36.98	47.06	57.00	64.90
MdSSUB	**2.20**	**2.53**	2.79	3.04	3.29	31.91	41.75	50.93	59.07	64.98
RISSUB	2.08	2.41	2.71	2.99	3.27	33.20	41.03	49.32	56.01	62.35
LOGAcMME	2.01	2.43	2.79	3.09	3.38	27.39	38.63	50.27	60.64	66.62
LOGMME	1.92	2.32	2.69	3.03	3.37	31.80	41.90	52.31	57.97	62.36
LOGRIMME	1.97	2.37	2.72	3.05	3.38	32.69	42.69	53.32	59.38	63.27
AcKal	2.14	2.51	**2.83**	**3.15**	3.43	**36.32**	**46.77**	**56.40**	**62.96**	**67.00**
MdKal	1.94	2.33	2.64	2.91	3.16	30.13	40.80	50.23	57.00	61.10

GRIMME, it was noted that these methods incorporated the use of a smoothing parameter α which provided a trade-off between musical type noise distortion and slurring in the resulting reconstructed speech. The value of α applied was determined experimentally using human listening tests. Therefore, a preliminary investigation was made to determine if this parameter might significantly impact on the resulting recognition rates. Results for experiments were conducted utilising various α values between 0.96 and 0.998 and the results are shown in Table 4. The recognition rates shown are for the first 110 utterances of the test corpus, with the stimuli being corrupted with 5 dB of babble noise. Results show that reducing the value of α improved recognition rates, particularly for LOGMME. This suggests that further improvement could be attained by further tuning of the modulation domain based algorithms so that they are optimised for speech recognition.

4. Conclusions

In this study, the ASR performance of modulation domain-based speech enhancement methods was compared with that of their acoustic domain counterparts, when used as a preprocessor of speech prior to the ASR feature extraction. The aim was to determine if recently reported modulation domain methods, which were tuned for human listening, would also offer additional advantages when coupled with an ASR system. In the ASR experiments performed on the TIMIT speech corpus, it was found that certain methods had performed differently and for different noises. For the spectral-subtraction algorithms, the modulation domain methods were found to be universally better than their acoustic-based ones in phoneme correctness for all noises. For the logMMSE-based methods, the acoustic domain variant was found to be more effective for all noises except for white noise. To explain why the modulation-domain MME methods were under-performing, preliminary tests were performed to determine if better tuning parameters could improve the ASR performance. The results demonstrated that further tuning have the potential to improve their competitiveness in ASR performance. Lastly, common speech enhancement metrics for speech quality and intelligibility were not found to be reliably indicative of ASR accuracy.

5. References

[1] K. Paliwal, J. Lyons, S. So, A. Stark, and K. Wójcicki, "Comparative evaluation of speech enhancement methods for robust automatic speech recognition," in *International Conference of Signal Processing and Communication Systems (ICSPCS)*. Gold Coast, Australia: IEEE, Dec 2010, pp. 1–5.

[2] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-27, no. 2, pp. 113–120, 1979.

[3] N. Wiener, *The Extrapolation, Interpolation, and Smoothing of Stationary Time Series with Engineering Applications*. New York: Wiley, 1949.

[4] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-32, no. 6, pp. 1109–1121, Dec 1984.

[5] ——, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-33, no. 2, pp. 443–445, Apr 1985.

[6] H. Hermansky and N. Morgan, "RASTA processing of speech," *IEEE Trans. Speech Audio Process.*, vol. 2, pp. 578–589, Oct 1994.

[7] Y. Zhang and Y. Zhao, "Spectral subtraction on real and imaginary modulation spectra," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Process. (ICASSP)*, Prague, May 2011, pp. 4744–4747.

Table 3: *TIMIT enhancement evaluation in terms of mean STOI scores and segmental SNRs for babble, F16, factory, and white noises (Highest scores are in bold).*

Algorithm	Mean STOI					Mean Segmental SNR				
SNR (dB)	0	5	10	15	20	0	5	10	15	20
Noisy (babble)	-	-	-	-	-	-6.871	-5.997	-5.395	-5.016	-4.778
AcSSUB	0.620	0.761	0.866	0.934	0.969	-3.058	1.090	5.147	9.294	13.530
MdSSUB	**0.670**	**0.793**	**0.884**	**0.944**	**0.975**	-3.436	0.784	4.839	8.946	13.174
RISSUB	0.667	0.792	0.883	0.941	0.971	-3.085	1.033	4.702	7.989	10.821
LOGAcMME	0.656	0.772	0.863	0.925	0.962	-4.015	0.216	4.284	8.350	12.429
LOGMME	0.643	0.763	0.862	0.933	0.971	**-1.954**	**1.951**	**5.868**	**9.963**	**14.103**
LOGRIMME	0.651	0.776	0.874	0.940	0.974	-2.226	1.785	5.691	9.573	13.255
AcKal	0.600	0.700	0.776	0.834	0.876	-7.252	-4.155	-1.643	0.415	2.128
MdKal	0.658	0.774	0.862	0.923	0.969	-3.306	0.998	4.947	8.571	11.797
Noisy (F16)	-	-	-	-	-	-6.934	-6.042	-5.423	-5.032	-4.788
AcSSUB	0.647	0.789	0.890	0.950	0.978	-2.281	1.898	5.954	10.086	14.203
MdSSUB	0.751	**0.856**	**0.924**	**0.964**	**0.984**	0.063	3.559	6.996	10.635	14.498
RISSUB	**0.754**	0.854	0.920	0.959	0.979	**0.821**	3.912	6.765	9.418	11.766
LOGAcMME	0.711	0.818	0.895	0.944	0.972	-1.573	2.266	5.962	9.658	13.485
LOGMME	0.687	0.800	0.893	0.951	0.979	-0.107	3.465	7.237	**11.013**	**14.866**
LOGRIMME	0.707	0.823	0.907	0.957	0.981	0.372	**3.922**	**7.429**	10.824	14.126
AcKal	0.729	0.817	0.882	0.926	0.950	-1.563	1.607	4.261	6.615	8.761
MdKal	0.704	0.811	0.889	0.938	0.968	-1.185	2.656	6.164	9.402	12.299
Noisy (white)	-	-	-	-	-	-6.981	-6.073	-5.441	-5.042	-4.795
AcSSUB	0.601	0.762	0.880	0.948	0.979	-2.449	1.788	5.874	9.971	**14.027**
MdSSUB	0.718	**0.833**	**0.910**	**0.955**	**0.980**	0.421	3.541	6.715	10.080	13.691
RISSUB	0.706	0.822	0.901	0.948	0.974	0.926	3.697	6.340	8.849	11.140
LOGAcMME	0.686	0.797	0.881	0.935	0.967	-1.136	2.379	5.798	9.268	12.886
LOGMME	0.668	0.792	0.887	0.941	0.972	**0.234**	3.587	6.942	**10.250**	13.743
LOGRIMME	0.679	0.804	0.893	0.944	0.973	0.488	**3.783**	**6.976**	10.058	13.156
AcKal	**0.743**	0.831	0.900	0.945	0.970	-0.030	3.142	6.201	9.207	12.035
MdKal	0.685	0.798	0.884	0.939	0.968	-1.070	2.680	6.115	9.299	12.126

Table 4: *Effect of α on the recognition rates of LogMME. Test evaluated using model generated from training set, and reduced test set (dr1) containing 110 utterances and input SNR of 5 dB of babble noise (Highest scores are in bold)*

α value	LogMME$_M$ Corr (%)	LogRIMME Corr (%)
0.998	40.92	41.94
0.996	43.52	42.73
0.990	44.42	44.98
0.980	44.98	44.76
0.970	**47.13**	**45.43**
0.960	46.67	45.55

[8] K. Paliwal, K. Wójcicki, and B. Schwerin, "Single-channel speech enhancement using spectral subtraction in the short-time modulation domain," *Speech Communication*, vol. 52, no. 5, pp. 450–475, May 2010.

[9] K. Paliwal, B. Schwerin, and K. Wójcicki, "Single channel speech enhancement using MMSE estimation of short-time modulation magnitude spectrum," in *Proc. ISCA Conf. Int. Speech Commun. Assoc. (INTERSPEECH)*, Florence, Italy, Aug 2011, pp. 1209–1212.

[10] S. So and K. Paliwal, "Modulation-domain Kalman filtering for single-channel speech enhancement," *Speech Communication*, vol. 53, no. 6, pp. 818–829, July 2011.

[11] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-28, no. 4, pp. 357–366, 1980.

[12] S. Furui, "Speaker-independent isolated word recognition using dynamic features of speech spectrum," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 34, no. 1, pp. 52–59, Feb 1986.

[13] C. Nadeu, P. Pachés-Leal, and B.-H. Juang, "Filtering the time sequences of spectral parameters for speech recognition,," *Speech Communication*, vol. 22, no. 4, pp. 315–332, Sep 1997.

[14] J. Garofolo, L. Lamel, W. Fisher, J. Fiscus, and D. Pallett, "DARPA TIMIT acoustic-phonetic continous speech corpus CD-ROM. NIST speech disc 1-1.1," *NASA STI/Recon Technical Report N*, vol. 93, 1993.

[15] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, *The HTK Book (for HTK Version 3.4)*, Engineering Department, Cambridge University, 2006.

[16] A. Varga and H. Steeneken, "Assessment for automatic speech recognition II: NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Communication*, vol. 12, no. 3, pp. 247–251, Jul 1993.

[17] P. Loizou, *Speech Enhancement: Theory and Practice*. Boca Raton, FL: Taylor and Francis, 2007.

[18] Y. Zhang and Y. Zhao, "Real and imaginary modulation spectral subtraction for speech enhancement," *Speech Communication*, vol. 55, no. 4, pp. 509–522, May 2013.

[19] B. Schwerin and K. Paliwal, "Using STFT real and imaginary parts of modulation signals for MMSE-based speech enhancement," *Speech Comm.*, vol. 58, pp. 49–68, Mar 2014.

[20] S. So, A. George, R. Ghosh, and K. Paliwal, "A non-iterative Kalman filtering algorithm with dynamic gain adjustment for single-channel speech enhancement," *International Journal of Signal Processing Systems*, vol. 4, no. 4, pp. 263–268, Aug 2016.

[21] K.-F. Lee and H.-W. Hon, "Speaker-independent phone recognition using hidden markov models," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 37, no. 11, pp. 1641–1648, Nov 1989.

[22] C. Taal, R. Hendriks, R. Heusdens, and J. Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Process. (ICASSP)*, Dallas, Texas, USA, Mar 2010, pp. 4214–4217.

A method for classifying voice quality

Adele Gregory

Department of Speech Pathology, James Cook University, Townsville, Australia
adele.gregory@jcu.edu.au

Abstract

The classification of voice quality is utilized across the fields of linguistics and speech pathology. This paper proposes a methodology to classify voice quality utilizing both auditory perceptual and acoustic references that provide a high level of inter-rater reliability. Using a case study from the field of infant language acquisition we show that this classification scheme provides a systematic way of combining spectrographic and wave inspection together with an auditory impression for a replicable methodology.

Index Terms: voice quality, classification, perceptual, acoustic

1. Introduction

Traditionally the quality or timbre of a sound has been defined as "that attribute of auditory sensation in terms of which a listener can judge that two sounds similarly presented and having the same loudness and pitch are dissimilar" [1: 45]. Given this broad definition, it is no surprise that previous reviews have reported as many as 67 terms for vocal quality in the literature [2]. Multiple systems such as those by [3, 4, 5] and [6] have endeavored to capture the variety of phonation types. These systems provide a wide range of options for classifying non-modal voice quality. They are routinely used in a number of fields, and auditory-perceptual evaluation of voice quality is the most commonly used clinical voice assessment method [7]. Perceptual methods are often utilized to classify consensus features of speakers during the production of sustained vowels, sentences, and running speech. Perceptual protocols are often referenced in the literature, and many studies have developed these, primarily for the evaluation of pathological voices. However perceptual evaluation has been heavily criticized because it is subjective [8]. [8: 14] notes that "building detail and complexity into a coding system does not guarantee that the listener's auditory-perceptual decision space is equal to the demands of the analysis task." As such a broader transcription that is more dependable has benefits over a narrow transcription with more detailed information but poorer reliability. In addition, perceptual methods do not necessarily provide information about actual vocal tract function; the complicated relationship between human auditory perception, acoustic measures of voice quality and vocal tract configuration limits the extent which perception of voice quality can reveal the underlying vocal tract physiology [7]. Despite this, a perceptual analysis allows for considerable advantages in terms of convenience, economy, and robustness [8]. Therefore [7] recommends multiple methods of voice quality evaluation: both subjective notation and appropriately implemented instrumental measures.

These instrumental measures may be aerodynamic or acoustic. Aerodynamic measures (e.g., subglottal pressure, phonation threshold pressure maximum flow declination rate) are used extensively in the clinical environment to diagnose and examine the differences pre- and post-treatment [10,11]. Acoustic measures (e.g., *f0*, waveform and spectral inspection, jitter and Long Term Average Spectrum) are also utilized in both diagnosis and treatment evaluation. Multiple measures are required, as voice quality does not have a single acoustic or aerodynamic correlate. Evidence suggests that voice quality may not even be independent of frequency and amplitude [9]. In addition to pitch and loudness, which are easily quantifiable as they have single acoustic correlates, voice quality is influenced by numerous different factors; including effects of the spectral envelope and its changes in time, periodic fluctuations of amplitude or fundamental frequency, and any noise component in the signal. Often measures are selected based on their suitability for a particular population. This paper presents a case study of how the consideration of these measures impact on methodology.

Infants form an interesting population for voice quality study as a large proportion of the vocalizations they produce are deemed to be non-modal [12]. This is a result of infants still developing the control and coordination necessary for the acquisition of speech. [12] created a regime classification system that utilized both perceptual and instrumental aspects of voice quality evaluation. However, they also introduced specific infant related phonation categories. In contrast [13, 14] created a system that utilized perceptual aspects and adult voice quality categories with infants. These two ideas can be combined so that a classification scheme is developed that can utilize common adult voice quality categories but will also take into account the perceptual and instrumental aspects of voice quality evaluation.

2. Methodology

A subset of 8 voice quality categories was chosen after consideration from a broad range of literature: harsh voice, creaky voice, whispery voice, modal voice, breathy voice, loft, whisper and voiceless. [13, 14, 15] have previously used these terms in an auditory analysis of voice quality parameters where the degree of laryngeal constriction (defined primarily in terms of the degree of sphincteringof the aryepiglottic fold in the larynx) distinguishes phonation types. These categories are based on laryngoscopic observations of the adult pharynx and larynx [16]. Thus there is some defined relationship between the vocal tract configuration and the perceptual voice quality categories in this proposed classification scheme. The use of these terms within this framework provides a broad spectrum of vocal behavior and an ability to speak to some extent of the vocal tract configuration. As such the methodology is similar to [12: 553] because "it [is] internally

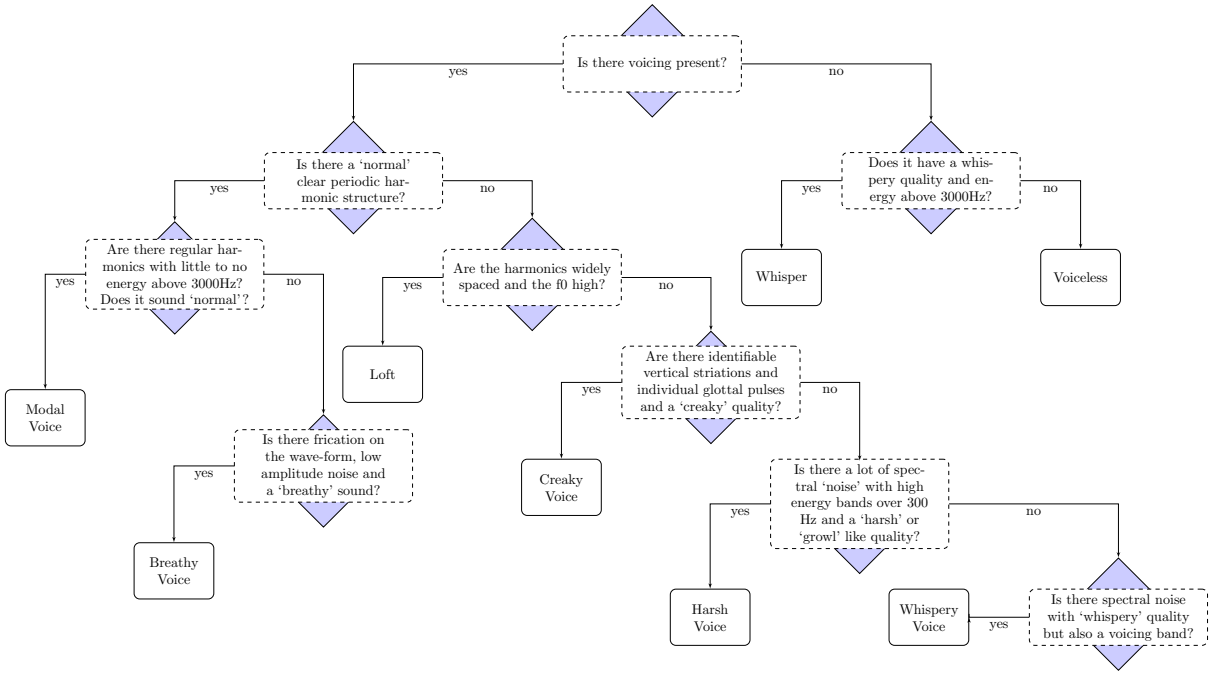


Figure 1*: Proposed classification schema for labelling voice quality categories*

comprehensive in the sense that all phonation are classified but it is not exhaustive in the sense that all vocal qualities are thereby represented." It thus takes into account [8]'s requirements for limiting complexity and detail. Acoustic correlates for each category were developed based on previous literature (e.g. [17], [18]). These included *f0*, waveform, and broadband spectrogram visual inspection. A decision–making tree was created based on these different categories (see Figure 1).

To test the methodology, a corpus of infant data was classified for voice quality. Two female infants comprised the study. Each infant was video-recorded using a Sony DCR TRV16E digital video camera recorder with an integrated electret condenser microphone. This integrated microphone was shown to have a wide-band flat response through the range of 100Hz to 10kHz. Infants were recorded in 16-bit stereo and at a sampling rate of 48kHz. Each recording session occurred in the infant's home in the presence and under the control of one of their caregivers. The camera was positioned on a stationary tripod and directed at the infant at a distance of approximately one to two meters. The infants were recorded for approximately one hour each week during the study (from 3 weeks up to 26 weeks of age). A full range of vocal behavior was recorded. Randomly selected vocalizations from each week of the study were chosen. For the purposes of the case study, a vocalization was any discrete sound produced with varying degrees of constriction occurring within one respiration cycle. Each vocalization was labeled by utilizing the classification tool by the author. A total of 761 separate vocalizations were labeled.

As auditory perceptual judgments are susceptible to a variety of sources of error and bias, inter-rater reliability measurements were used to quantify the consistency with which these judgments are made. An independent analyst was trained to utilize the classification scheme on practice tokens for approximately an hour. Any reflexive sounds (hiccup, sneeze, and cough) were labeled as such; all other vocalizations were labeled for voice quality. If the independent analyst was unsure about the category, they made the decision based on their perception of the sound irrespective of the acoustic references. The tokens were presented to them in a randomized order, and they were blinded to the participant.

3. Results

A total of 761 vocalizations were labeled with a voice quality category by two analysts. The categories were then compared. The scheme can provide a technique for the comprehensive classification of a child's recorded productions, but in this case study, 96% of all vocalizations were analyzed by both raters as 4% were left unlabeled by the independent analyst. No attempts were made to establish a consensus or to re-train the independent analyst for increased discrimination. Table 1 provides the results from the study. Overall 82% or Cohen's Kappa of 0.77 utterances were labeled independently in agreement representing a high level of agreement. Modal, creaky, harsh, breathy, loft and voiceless voice qualities were all recognized with accuracies approaching or exceeding 80%. When the errors for each voice quality were examined in turn the following results emerge:

- Creaky voice: most confused with harsh voice
  - (n=8, 7%)
- Harsh voice: most confused with creaky voice
  - (n = 4, 4%)
- Whispery voice: most confused with breathy voice
  - (n = 3, 13%)
- Modal voice: most confused with breathy voice
  - (n = 10, 6%)
- Breathy voice: most confused with modal voice
  - (n = 4, 8%)
- Loft: most confused with modal voice
  - (n = 2, 6%)
- Whisper: confused with breathy voice and voiceless
  - (n =1, 13%)

Analyst 2 (Independent)	Analyst 1: Creaky	Harsh	Whispery Voice	Modal	Breathy	Loft	Whisper	Voiceless	Hiccup	Cough	Sneeze	Unlabeled	Totals
Creaky	99	4		3		1		1					108
Harsh	8	84		8		1		5					106
Whispery Voice	0	3	18	4	3								28
Modal	6	3	2	140	4	2		2					159
Breathy	1	3	3	10	42	1	1						61
Loft		3		1		28							32
Whisper							6	3					9
Voiceless		1		7	1		1	197					207
Hiccup						1		3	4				8
Cough		3			1			5	1		1		11
Sneeze								2			3		5
Unlabeled	2	1		4				20					27
Totals	116	105	23	177	51	34	8	238	5	0	4	0	761

Table 1*: Results from analysts' determining the voice quality of each vocalization*

- Voiceless: either labeled as vegetative vocalisation or not labeled
    - (n = 10, 4% and n = 20, 8%)

Whilst the Kappa statistics indicated overall good performance, it is evident that some voice qualities were either more challenging to classify accurately or with confidence. This was particularly true for voiceless vocalizations where the largest number (n=20) of unlabeled vocalization was located. As the classification of voice quality categories required definitive decisions the first type of error was in determining phonation. This occurs at the first node ('Is there voicing present?') (see Figure 2). There was the possibility of an analyst recording a voice quality where there was none (Category 1a), or saying there was none when there was eligible phonation (Category 1b). The largest number of Category 1a errors was where voiceless vocalizations were labeled with the combined vegetative class (cough, sneeze, hiccup, n=10) and where the analyst was unsure and therefore did not make a decision (unlabeled n=20). However, when these are removed, there was a high level of agreement about whether a vocalization was voiceless or had some sort of phonation present (94%). The largest number of Category 1b errors was when the independent analyst misjudged a modal voice as voiceless (n=7).

Other voice qualities were also confused for each other. Category 2 errors account for the misclassification between modal and breathy. Category 3 errors account for the misclassification between creaky and harsh. These category errors occur because the scheme requires a forced decision.

It is also suggested that voice quality categories may be incorrectly categorized due to the perceptual similarities between them (e.g. whispery voice and breathy voice). Improved familiarity with borderline cases would help differentiate between the more easily confused voice qualities. Future training of analysts should focus on this. Qualitative review questions about the ease and use of the scheme and underlying issues of classification would also add to the understanding about why some voice qualities cannot be categorized confidently.

4. Discussion

The results from this case study have shown that infant vocalizations can be exhaustively classified by the scheme shown at Figure 1. The combination of acoustic and perceptual cues to decide on the voice quality laid out in the decision tree enables a systematic way to delineate voice quality in infant vocalizations. The results of the inter-rater reliability study compare very favorably with those of [12]. When categorizing infant voice quality using their schema, [12: 560] obtained a Cohen's Kappa of 0.76. The Kappa statistics indicate overall good performance. To enhance the scheme amendments have been made as shown in Figure 2. These minor changes remove two unnecessary decision nodes.

By demonstrating this scheme's utility on an area of linguistic study known for its challenges it can be inferred that it will fare as well or better in other areas of study. Further application of this methodology is not limited to the field of language acquisition. As defining voice quality categories across disciplines continues to be an issue (including inter-rater and inter-lab reliability) this scheme may fulfill a need for a replicable method of voice quality classification. Due to the voice quality categories being based on adults rather than infants it would be possible to extend the scheme to work with adult populations as well. This scheme particularly commends itself for application in any field where the classification of voice quality provides discriminatory insight including but not limited to the fields of socio-linguistics, forensic linguistics and speech pathology. Preliminary studies are currently underway investigating its suitability for its use with

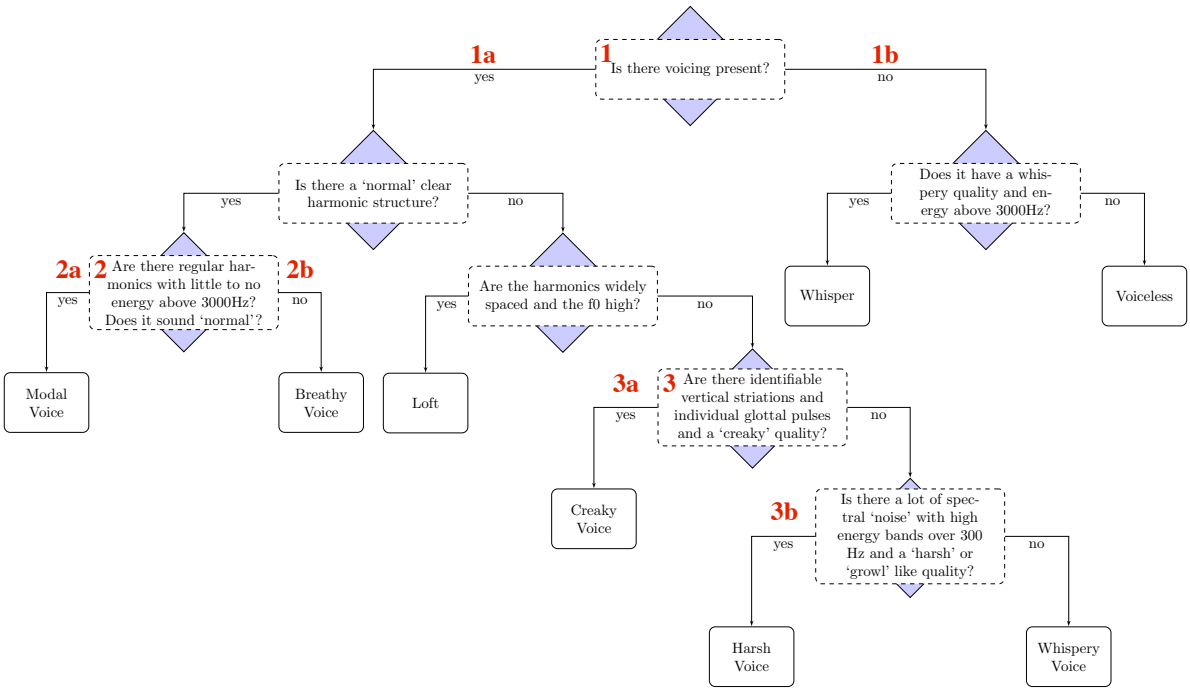


Figure 2: *Amended classification schema for labelling voice quality categories*

disordered voices.

Within the field of infant language acquisition, the technique provides the ability to show the change in dominant voice qualities over time as well as adding other acoustic dimensions (e.g., *f0* or duration) to the task of describing speech development in these early stages [19]. Extended longitudinal research with large sets of infants including from a variety of linguistic backgrounds would provide additional data to systematically delineate the development of vocal quality and the emergence of control over modal voice.

5. Conclusions

Voice quality is variously utilized across the fields of linguistics and speech pathology. This paper proposes a methodology to classify voice quality utilizing both auditory perceptual and acoustic references that provide a high level of inter-rater reliability. Using a case study from the field of infant language acquisition, it is shown that this classification scheme provides a systematic way of combining spectrographic and wave inspection together with an auditory impression for a replicable methodology.

6. References

[1] ANSI. *USA standard: Acoustical terminology (S1.1)*. American National Standards Institute, New York, 1960.

[2] Pannbacker, M. Classification Systems of Voice Disorders: A review of the literature. *Language, Speech & Hearing Services in Schools*, 15:169–179, 1984.

[3] Laver, J. *The phonetic description of voice quality*. Cambridge University Press, Cambridge, 1980.

[4]Laver, J. *Principles of Phonetics*. Cambridge University Press, Cambridge, 1994.

[5] Laver, J. Phonetic Evaluation of Voice Quality. In Kent, R. D. and Ball, M. J., editors, *Voice quality measurement*, pages 37– 48. Singular Publishing Group, San Deigo, 2000.

[6] Ball, M. J., Esling, J. H., and Dickson, C. The Transcription of Voice Quality. In Kent, R. D. and Ball, M. J., editors, *Voice quality measurement*, pages 49–58. Singular Publishing Group, San Deigo, 2000.

[7] Oates, J. Auditory-perceptual evaluation of disordered voice quality. *Folia Phoniatrica et Logopaedica*, 61(1):49–56, 2009.

[8] Kent, R. D. Hearing and believing: Some limits to the auditory-perceptual assessment of speech and voice disorders. *American Journal of Speech-Language Pathology*, 5(3):7–23, 1996.

[9] Krumhansl, C. L. and Iverson, P. Perceptual interactions between musical pitch and timbre. *Journal of Experimental Psychology: Human Perception and Performance*, 18(3):739–751, 1992.

[10] Hartl, D. M., Hans, S., Vaissière, J., and Brasnu, D. F. Objective acoustic and aerodynamic measures of breathiness in paralytic dysphonia. *European Archives of Oto-rhino-laryngology*, 260(4):175–182, 2003.

[11]Giovanni, A., Revis, J., and Triglia, J.-M. Objective Aerodynamic and Acoustic Measurement of Voice Improvement After Phonosurgery. *The Laryngoscope*, 109 (4):656–660, 1999.

[12] Buder, E. H., Chorna, L. B., Oller, D. K., and Robinson, R. B. Vibratory Regime Classification of Infant Phonation. *Journal of Voice*, 22(5):553–564, 2008.

[13] Esling, J. H. and Harris, J. G. States of the Glottis: An Articulatory Phonetic Model Based on Laryngoscopic Observations. In Hardcastle, W. J. and Beck, J. M., editors, *A Figure of Speech*, pages 347–353. Lawrence Erlbaum Associates, Mahwah, 2005.

[14] Benner, A., Grenon, I., and Esling, J. H. Infants' phonetic acquisition of voice quality parameters in the first year of life. In *16th International Congress of Phonetic Sciences*, 2073–2076, Saarbrücken, 2007.

[15] Esling, J. H. There are no back vowels: The laryngeal articulator model. *The Canadian journal of linguistics/La revue canadienne de linguistique*, 50(1):13–44, 2005.

[16] Edmondson, J. A. and Esling, J. H. The valves of the throat and their functioning in tone, vocal register and stress: laryngoscopic case studies. *Phonology*, 23(02): 157–191, 2006.

[17] Ladefoged, P. *Phonetic data analysis: An introduction to fieldwork and instrumental techniques.* Blackwell Publishing Melbourne, 2003.

[18] Keating, P. A. and Esposito, C.M. Linguistic voice quality. *Working papers in Phonetics,* 1-8, 2007.

[19] Gregory, A., Tabain, M. and Robb, M. Duration and voice quality of early infant vocalizations. *Journal of Speech, Language and Hearing Research,* [advance online publication]; 1-12, 2018.

Japanese Vowel Devoicing Modulates Perceptual Epenthesis

Alexander J. Kilpatrick[1], Shigeto Kawahara[2], Rikke L. Bundgaard-Nielsen[3], Brett J. Baker[1], Janet Fletcher[1]

[1]University of Melbourne
[2]Keio University
[3]MARCS Institute for Brain, Behaviour and Development, Western Sydney University
alex.kilpatrick@unimelb.edu.au

Abstract

This study investigates a relationship between perceptual epenthesis and vowel devoicing in Japanese. Across two experiments, epenthetic vowels are compared in environments where devoicing and deletion occur. In Experiment 1, participants assign illicit /VCCV/ and /VCVC/ tokens to /VCuCV/ and /VCVCu/ categories and judge how well tokens fit to the allocated category. In Experiment 2, participants discriminate between phonotactically illicit and licit tokens in AXB tests. The results show that illicit tokens are a better match to—and more difficult to discriminate from—their perceptually nearest legal counterpart when the target vowels are deleted than when they are merely devoiced.

Index Terms: Phonology, Phonetics, Japanese, Perceptual Epenthesis, Vowel Devoicing, Perceptual Assimilation.

1. Introduction

Standard Japanese (hereafter: Japanese) phonotactics do not allow non-homorganic consonant clusters and word-final, non-nasal consonants. As a result, Japanese listeners sometimes perceive an illusory, epenthetic /u/, which serves to perceptually repair the input to adhere to Japanese phonotactics [1], when they are exposed to such violations. It has been proposed that /u/ is epenthesised in these contexts because it is the shortest of all Japanese vowels [2], making it the phonetically minimal element of the language [1]. In line with a novel extension of the Perceptual Assimilation model [3], which predicts and accounts for the influence of L1 transitional probability on L2 perception [4], we propose that perceptual epenthesis is a process whereby illicit non-homorganic consonant clusters and word-final, non-nasal consonants are assimilated to their perceptually nearest and most predictable match. This process of assimilation reduces—or even eliminates—the perceptual distance between tokens that contain ether illicit consonant clusters (VCCV) or word-final, non-nasal consonants (VCVC) and tokens that adhere to Japanese phonotactics (VCVCV), resulting in the illusory vowel effect.

Vowel devoicing in Japanese can occur with all phonemically short vowels (hereafter: vowels); however, it is only systematic with high vowels (/u/ and /i/) [5]. Japanese vowel devoicing typically occurs either between voiceless consonants (/C̥VC̥/) or after a voiceless consonant at the end of a word (/C̥V#/) [6]. A comprehensive corpus analysis identifies several other contributing factors that influence vowel devoicing including the manner of articulation of the preceding (C_1) and following (C_2) consonant, as well as variation among individual phones [7]. For example, when /u/ occurs between two voiceless consonants, it undergoes devoicing 84% of the time, this increases to 98% if the C_1 is a voiceless fricative and the C_2 is a voiceless plosive, and further still to 99% if the C_1 is an /s/ and the C_2 is a /p/ (See Table 1). Others have proposed that the predictability of the preceding consonant in CV sequences [8] or the frequency of the words that carry the target vowel [9] can affect deletion/devoicing.

In Japanese, devoiced vowels that follow stops are typically phonetically realized differently from those that follow fricatives and affricates. Some argue that devoiced vowels that follow fricatives and affricates are deleted (See discussion in [6]) and that these instances of deletion may result in consonantal syllables [10]. Electropalatography [10] and electromagnetic articulography [11] studies have also suggested that these vowels are undergoing deletion, with speakers not exhibiting lingual configurations typical of the Japanese /u/ in some devoiced tokens.

In the following, we refer to these post-fricative devoiced allophones as “deleted” vowels in order to distinguish them from devoiced vowels that follow voiceless stops. Phonological vowel deletion in these contexts would mean that voiceless non-homorganic consonant clusters were phonotactically permissible in Japanese and therefore unlikely to elicit perceptual epenthesis; numerous experiments—including those featured in the present paper—have shown that this is not the case. Instead, we propose that some surface level representation of deleted vowels remain but that these near-zero allophones act as a better match to vowel-less sequences. These near-zero allophones maintain fewer or less salient acoustic cues than devoiced or voiced vowels, making them perceptually minimal.

In line with the aforementioned extension of the Perceptual Assimilation model, we also propose that the perceptual minimality of the expected allophone has an influence on the perceptual distance between illicit sequences and their perceptually nearest, predictable match. Indeed, we argue that this is due to sequences that predictably stimulate perceptually minimal allophones eliciting less discriminable illusory vowels because the assimilation distance between the illicit sequence and its nearest match is narrowed. We test this hypothesis in two experiments that examine epenthetic vowels which occur after voiceless fricatives, voiceless plosives and voiced consonants. In Experiment 1, participants assign VCCV and VCVC tokens to VCVCV categories and assign goodness of fit (GoF) ratings to how well they adhere to a given category. In Experiment 2, participants discriminate VCCV and VCVC tokens from VCVCV tokens in a series of AXB discrimination experiments.

.

Table 1. *List of experimental tokens and rates at which /u/ undergoes devoicing in Japanese discourse. Here, Env .% = devoicing rates in the specific environment tested, Manner = rates based on the manner of articulation of the C_1 and C_2 in /ÇVÇ/ sequences and Voicing = devoicing rates based purely on the voice/voicelessness of the C_1 and C_2 [7].*

Location	Allophone	Licit Token	Illicit Token	Environment	Env. %	Manner	Voicing
Medial	Deleted	/esupo/	/espo/	/s_p/	99%	98%	84%
Medial	Devoiced	/ekupo/	/ekpo/	/k_p/	88%	80%	84%
Medial	Devoiced	/epuso/	/epso/	/p_s/	60%	74%	84%
Medial	Voiced	/egupo/	/egpo/	/g_p/	N/A	N/A	2%
Medial	Voiced	/ezubo/	/ezbo/	/z_b/	N/A	N/A	1%
Medial	Voiced	/ebuzo/	/ebzo/	/b_z/	N/A	N/A	1%
Final	Deleted	/epusu/	/epus/	/s_#/	N/A	N/A	N/A
Final	Devoiced	/esupu/	/esup/	/p_#/	N/A	N/A	N/A

2. Method

2.1. Stimuli

A full list of all 16 tokens appears in Table 1. The stimuli were produced by three phonetically trained female Australian English (AustE) speakers. These were recorded in a recording studio located at the University of Melbourne and had a bit depth of 64kb/sec and a sample rate of 48kHz. Each speaker produced five consecutive repetitions of each of the 16 tokens. The first and fifth repetitions were not used in Experiment 1 to avoid any effects of list initial unfamiliarity and list final intonation patterns. The remaining three tokens were excised with a 20 ms ramp-in and a 10 ms ramp-out. On average, /u/ duration in medial licit tokens was 85 ms (range 67-106 ms, *SD* =13 ms); average target /u/ duration in word-final licit tokens was 152 ms (range 95-279 ms, *SD* = 40 ms). Contrasting licit/illicit token pairs were designed so that the production of licit stimuli would predictably produce varying allophones in the target /u/; deleted (e.g., /es**u**po/), devoiced (e.g., ep**u**so) and voiced (e.g., /ez**u**bo/) (see Table 2).

2.2. Participants

34 undergraduate students from the Mita campus of Keio University were recruited as participants for Experiments 1 and 2. Participants were all L1 Japanese speakers, born to L1 Japanese speaking parents. Participants were recruited by word of mouth. Participant ages ranged from 18 to 26 (*M* = 20, *SD* = 1.6) and were selected on the basis of limited exposure to languages other than Japanese although all participants had previously studied English due to it being a compulsory subject in the Japanese education system.

2.3. Procedure: Experiment 1

Experiment 1 took place in a quiet room located at the Mita campus of Keio University. The experiment consisted of a single block of trials which contained all 16 tokens. Participants were asked to categorise tokens into 8 categories. These categories were presented to the participants as on-screen buttons with Hiragana labels (categories, tokens and Hiragana labels presented in Table 3). Tokens were drawn at random from a library of 144 stimuli (16 tokens x 3 speakers x 3 repetitions each). Upon assigning each token to a category, participants were asked to assign a GoF rating to indicate how well the token fit to the assigned category. This was presented to participants as a Likert scale ranging from 1 to 7. To explain that a low score was supposed to indicate a poor fit, the 1 on the Likert scale was labelled 違う (different) and the 7 was labelled 同じ (identical). We predict that listeners will assign higher GoF ratings to illicit tokens with C_1 voiceless fricatives due to deleted vowels being a better match to vowel-less sequences.

2.4. Procedure: Experiment 2

Experiment 2 was conducted directly after Experiment 1 in the same location. In Experiment 2, participants were required to respond to 192 AXB discrimination trials, 24 triads for each of the 8 licit/illicit contrasts (Table 2). To avoid speaker or phone sequence bias, tokens were organized into six speaker sequences (123, 132, 213, 231, 312, 321) and each of the speaker sequences was organised into four token sequences (AAB, ABB, BAA, BBA). All contrasts were presented to participants in a single block from which each AXB triads were drawn at random with a replace paradigm so that any trial that timed out was replayed later during the experiment. Both discrimination accuracy and response times were recorded. Tokens were spaced with a 1000 ms inter-stimulus interval. Here we predict that listeners will have greater difficulty discriminating between contrasts with voiceless fricatives in the C_1 position due to the smaller perceptual distance between the deleted vowel and the vowel-less sequence.

3. Results: Experiment 1

3.1. Categorisation Rate

Participants categorised most tokens to their perceptually nearest phonotactically licit category where the phonotactic violation is repaired by a /u/. All illicit tokens but one were categorised according to this prediction at a rate of 90% or greater. The one illicit token that did not adhere to a 90% categorisation rate was /egpo/, which was categorised as /ekupo/ 25% of the time. Licit tokens were also assigned to their predicted category at a rate of 90% or greater except in the case of the /epusu/ token which was categorised as /epuso/ 30% of the time and was only assigned to its predicted /epusu/ category 67% of the time; this is less than its illicit counterpart, /epus/, which was categorised as /epusu/ 95% of the time.

Table 2. *AXB contrasts organized by word position and the most likely target allophone in licit tokens.*

.	Deleted	Devoiced	Voiced
Word Medial	/esupo/-/espo/	/epuso/-/epso/	/ezubo/-/ezbo/
		/ekupo/-/ekpo/	/ebuzo/-/ebzo/
			/egupo/-/egpo/
Word Final	/epusu/-/epus/	/esupu/-/esup/	

Table 3. *Categorisation rates and Goodness of Fit ratings for licit and illicit tokens. Goodness of Fit ratings are presented in parenthesis. Categorisation rates less than 1% are not featured.*

	Medial Contrasts						Final Contrasts	
	えすぽ	えくぽ	えぷそ	えぐぽ	えずぼ	えぶぞ	えすぷ	えぷす
	esupo	ekupo	epuso	egupo	ezubo	ebuzo	esupu	epusu
/esupo/	98% (5.67)				1% (2.5)			1% (2.75)
/espo/	94% (5.81)				1% (2)			5% (5)
/ekupo/		90% (5.44)		10% (4.43)				
/ekpo/		91% (5.18)		9% (3.59)				
/epuso/			91% (5.5)			5% (2.93)	4% (3.83)	
/epso/			95% (5.18)			1% (2.75)	4% (3.17)	
/egupo/				100% (6.03)				
/egpo/		25% (4.13)		75% (4.79)				
/ezubo/					100% (5.43)			
/ezbo/	3% (4.5)				97% (4.62)			
/ebuzo/						100% (5.72)		
/ebzo/						100% (5.21)		
/epusu/			30% (4.47)			3% (2.56)	67% (4.62)	
/epus/			4% (3.00)			1% (2.75)	95% (4.78)	
/esupu/	9% (4.86)							90% (4.97)
/esup/	9% (1.93)							91% (3.81)

3.2. Goodness of Fit Rating

Overall, licit tokens achieved a higher average GoF rating (5.37) than illicit tokens (4.91); $t(271) = 6.24$, $p = < 0.001$. The only tokens that did not adhere to this pattern were those with a voiceless fricative preceding the epenthetic context. In these deleted contexts, the illicit tokens achieved higher GoF ratings than the licit tokens, /espo/ (5.81) was rated significantly higher than /esupo/ (5.67); $t(33) = -2.95$, $p = 0.006$, and /epus/ (4.78) was rated significantly higher than /epusu/ (4.62); $t(33) = -6.56$, $p = < 0.001$, despite being assigned to the /esupo/ and /epusu/ categories respectively.

Table 4. *Average scores for licit and illicit tokens, difference between scores and results from paired sample t-tests.*

Licit		**Illicit**				
Token	**GoF**	**Token**	**GoF**	**Diff.**	***t***	***p***
esupo	5.67	espo	5.81	-0.14	-6.87	< 0.001
ekupo	5.44	ekpo	5.18	0.26	10.92	< 0.001
epuso	5.5	epso	5.18	0.32	10.11	< 0.001
egupo	6.03	egpo	4.79	1.24	24.54	< 0.001
ezubo	5.43	ezbo	4.62	0.81	25.54	< 0.001
ebuzo	5.72	ebzo	5.21	0.51	17.29	< 0.001
epusu	4.62	epus	4.78	-0.16	-5.32	< 0.001
esupu	4.97	esup	3.81	1.16	31.11	< 0.001
Average	5.42		4.92	0.5		

4. Results: Experiment 2

4.1. Medial Contrasts

The discrimination accuracy results support our hypothesis that deletion contexts (e.g., /es**u**po/) are harder to discriminate from vowel-less tokens than devoiced or voiced contexts (e.g., /ep**u**so/ or /ez**u**bo/). Of the medial AXB tests, participants were least accurate at discriminating between deleted contrasts (M = 68%, SD = 14%), followed by devoiced contrasts (M = 75%, SD = 16%) and finally voiced contrasts (76%, SD = 14%). A one-way ANOVA between voicing conditions was conducted to compare the effect of voicing of the predictable allophone in the target position on test accuracy. The ANOVA revealed a significant main effect; $F(2, 201) = 3.1$, $p < 0.05$. Post hoc comparisons using the Bonferroni correction revealed a significant difference between deleted and voiced contrasts ($p < 0.05$) but not between deleted and devoiced ($p = 0.185$) or devoiced and voiced contrasts ($p = 1$).

In medial contrasts, response time results largely mirror the results in terms of accuracy whereby participants required more time to respond to contrasts that were difficult to discriminate. Participants took longest to respond to deleted contexts (M = 1279 ms, SD = 136 ms), followed by devoiced contexts (M = 1261 ms, SD = 135 ms), and finally voiced contexts (M = 1241 ms, SD = 139 ms). A one-way ANOVA of voicing conditions on response time also revealed a significant effect; $F(2, 4895) = 6.5$, $p < 0.01$. As with accuracy, a post hoc comparison with Bonferroni correction revealed a significant difference between deleted and voiced conditions ($p < 0.01$) but not between deleted and devoiced ($p = 0.087$) or devoiced and voiced ($p = 0.338$).

4.2. Word-Final Contrasts

As with medial contrasts, participants were less accurate at discriminating between the word-final deleted contrast (/epusu/-/epus/ M = 81%, SD = 8%) compared to the word-final devoiced contrast (/esupu/-/esup/ M = 86%, SD = 12%). A paired-samples t-test was conducted to compare the accuracy results of /epusu/-/epus/ and /esupu/-/esup/ contrasts. This revealed a significant difference between the two word-final contrasts ($t(33) = 2.5$, $p < 0.05$). Participants also took longer to respond to the /epusu/-/epus/ contrast (M = 1325 ms, SD = 131 ms) compared to the /esupu/-/esup/ contrast (M = 1299 ms, SD = 156 ms). A paired samples t-test calculated on this difference revealed a significant difference ($t(813) = 2.19$, $p < 0.05$).

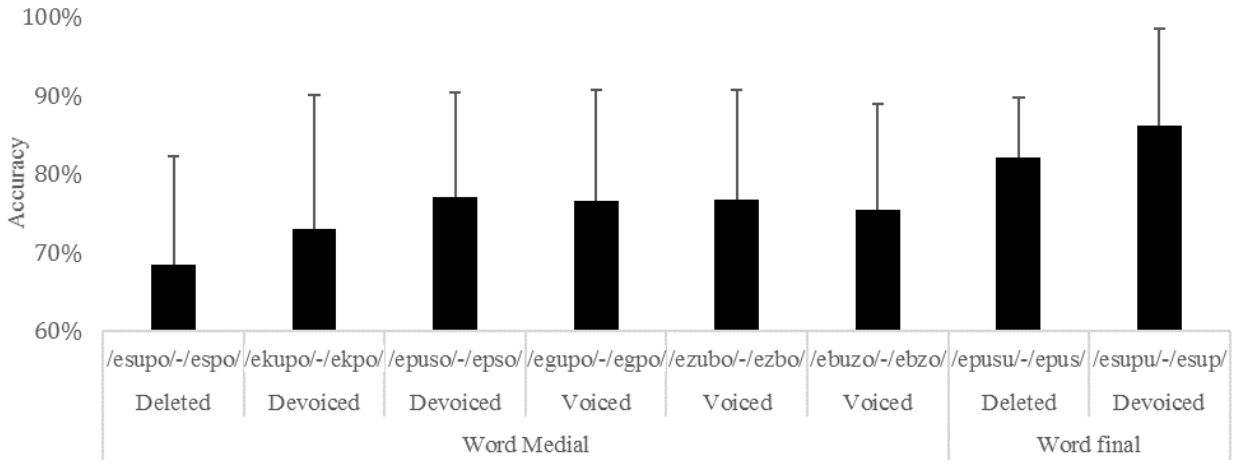


Figure 1. *AXB test accuracy. Error bars represent standard deviation.*

5. Discussion

5.1. Experiment 1: Categorisation and Goodness of Fit

In Experiment 1, we predicted that participants would assign higher GoF ratings to tokens with a voiceless fricative in the C_1 position. This hypothesis is supported by the results of this experiment whereby /espo/ and /epus/ tokens achieved higher GoF ratings than even their licit counterparts. We propose that this is likely due to the perceptual minimality of deleted vowels. Sequences that predictably elicit perceptually minimal vowels are a better match to vowel-less sequences than the voiced vowels produced by our AustE speaking volunteers.

5.2. Experiment 2: AXB Discrimination

In Experiment 2, we predicted that listeners would have more difficulty discriminating between contrasts where the epenthetic context would predictably undergo vowel deletion. This hypothesis is reflected in both medial and word-final discrimination accuracy results which show that contrasts were less discriminable when the C_1 was a voiceless fricative compared to other consonants. Deleted contrasts were significantly more difficult to discriminate than devoiced or voiced contrasts. This suggests that contrasts are more discriminable when the epenthetic vowel would predictably elicit voicing if the token were spoken by a Japanese speaker.

6. Conclusion

The present report demonstrates that Japanese listeners are more likely to perceive an epenthetic /u/ when the C_1 is a voiceless fricative when compared with voiceless stops or voiced consonants. In line with the aforementioned extension of PAM [4], phonotactically unattested sequences are assimilated to a predictable match. When the epenthetic context is preceded by a voiceless fricative, the assimilation distance is shortened due to the perceptual minimality of "deleted" vowels. One possible explanation for the difference between C_1 voiceless fricative and C_1 voiceless plosive contexts is that the turbulent aperiodic energy of the fricative masks the acoustic cues of the target vowel more substantially than the release of the stop. This masking makes these near-zero allophones a better match to vowel-less sequences, encouraging assimilation to the target phoneme. This assimilation reduces or eliminates the perceptual distance between illicit sequences and their nearest, most predictable match, making illicit tokens (e.g., /espo/) more acceptable and making contrast pairs (e.g., /espo/-/esupo/) less divergent.

7. Acknowledgements

We would like to thank our research participants and Keio University. We also thank the W.T. Mollison scholarship committee at the University of Melbourne, as well as Cathleen Benevanto, Rosey Billington, Katie Jepson and Eleanor Lewis.

8. References

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

[2] Arai, T., Warner, N., & Greenberg, S. "OGI tagengo denwa onsei koopasu-ni okeru nihongo shizen hatsuwa onsei no bunseki" [analysis of spontaneous Japanese in OGO multi-langauge telephone speech corpus], The Spring Meeting of the Acoustical Society of Japan, 1, 361-362, 2001.

[3] Best, C. T. "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, 1994.

[4] Kilpatrick, A. J., Bundgaard-Nielsen, R. L., & Baker, B. J. "Japanese Co-occurrence Restrictions Influence Second Language Perception", Applied Psycholinguistics, In Press.

[5] Maekawa, K. "Hatsuwa sokudo ni yoru yūsei kukan no hendō" [Effect of voicing rate on voicing variation in Japanese]. IEICE Technical Report SP89-148. 47–53, 1990.

[6] Fujimoto, M. "Vowel Devoicing", in H. Kubozono [Ed], Handbook of Japanese Phonetics and Phonology, Walter de Gruyter, 167-214, 2015.

[7] Maekawa, K., & Kikuchi, H. "Corpus-based analysis of vowel devoicing in spontaneous Japanese: an interim report.", in J. van de Weijer, K. Nanjo & T. Nishihara [Eds], Voicing in Japanese, 84, 205-228, 2005.

[8] Whang, J. "Recoverability-driven coarticulation: Acoustic evidence from Japanese high vowel devoicing." The Journal of the Acoustical Society of America, 143, 1159, 2018.

[9] Kilpatrick, A. J., Bundgaard-Nielsen, R. L., & Baker, B. J. "Japanese Vowel Deletion Occurs in Words in Citation Form", Proceedings of the 16th Australasian International Conference on Speech Science and Technology, 325-328, 2016.

[10] Matsui, M. "On the input information of the C/D model for vowel devoicing in Japanese." Journal of the Phonetic Society of Japan 21:1, 127-140, 2017.

[11] Shaw, J., & Kawahara, S. "The lingual gesture of devoiced /u/ in Tokyo Japanese." Journal of Phonetics, 66, 100-118, 2018.

Classification of interrogatives as information-seeking or rhetorical questions

Bettina Braun, Daniela Wochner, Katharina Zahner, Nicole Dehé

Department of Linguistics, University of Konstanz, Germany
{bettina.braun, daniela.wochner, katharina.zahner, nicole.dehe}@uni-konstanz.de

Abstract

Rhetorical questions (RQs) differ from information-seeking questions (ISQs) in that they do not require an answer and instead seek commitment of the addressee regarding the underlying proposition. We tested the prosodic differences between ISQs and RQs in German and showed that polar RQs were mainly realized with a high plateau (H-% in the GToBI system) and polar ISQs with a high-rise (H-^H%). *Wh*-RQs almost exclusively ended in a low edge tone whereas *wh*-ISQs showed more tonal variation (L-%, L-H%, H-^H%). Irrespective of question type, RQs were mainly produced with L*+H accents. Phonetically, RQs were – compared to ISQs – more often realized with breathy voice quality in the beginning of the utterance and with longer constituent durations. The object noun was particularly lengthened. Classification and regression trees showed that polar questions can be classified on the basis of the boundary tone alone, and *wh*-questions by an interaction between phonological events and duration. We discuss the findings with respect to the phonology-semantics interface.
Index Terms: rhetorical question, information-seeking question, prosody, classification, German

1. Introduction

This paper deals with the prosody of rhetorical and information-seeking questions in German in terms of intonational events and phonetic realization, and addresses the question of which of the parameters are needed to automatically classify utterances with an interrogative form as either rhetorical or information-seeking. RQs can have the form of a polar question (with subject-verb inversion in German, see (1)), or a *wh*-question with a fronted *wh*-element (see (2)). Other question types are also possible, but here we only investigate polar questions and *wh*-questions.

(1) Mag denn jemand Limonen?
Likes PRT anyone limes?
'Does anyone like limes?'

(2) Wer mag denn Limonen?
Who likes PRT limes?
'Who likes limes?'

The literature discusses RQs mostly in terms of their semantic and pragmatic properties: Canonical ISQs are used to seek information from the addressee. The answer to an ISQ can only be given by the addressee and not by the speaker [1]. In contrast, RQs do not require or expect answers from the addressee as stated by several authors [2-8]. Instead, they are used when the answer is obvious or at least inferable to all discourse participants [1, 9-11]. Moreover, the purpose of RQs is to seek the addressee's commitment to the proposition that is presupposed by the question [7]. Other functions of RQs are to change a topic or to engage the audience in monologues or retorts, (e.g., *Is the Pope Catholic*?, cf. [10], [12]). These other functions are not investigated in this paper. Regarding syntactic form and lexical cues, a rhetorical illocution may be signaled by strong negative polarity items (e.g., *Who on earth needs holidays*?, cf. [13], [7]), and, in German, by the discourse particles *schon* and *auch*[1], cf. [14], [7]. These discourse particles are sufficient to trigger an RQ interpretation, but they are not necessary. That is, RQs and ISQs can be string-identical on the surface and can be disambiguated by the context (3), as well as by their prosodic realization.

(3) ISQ context:
At a party, you offer cake made with limes. You would like to know which of the guests like this fruit and would like some of it. You say to your guests:
Q: Does anyone like limes?

RQ context:
Your aunt offers limes to her guests. However, it is known that this fruit is too sour to be eaten on its own. You say to your cousin:
Q: Does anyone like limes?

Previous pilot data from German [15] showed that polar RQs have a higher proportion of high plateaus (H-% in the GToBI annotation system [22]) than polar ISQs, which were typically produced with a high rise (H-^H%). *Wh*-questions generally ended in a fall (L-%), with a higher proportion of L*+H nuclear accents in *wh*-RQs than in *wh*-ISQs. RQs were also produced with longer constituent durations than ISQs and had a breathier voice quality. Here we present results from a more controlled production experiment. Based on the pilot data and on claims made in the literature on English [6, 8, 16, 17], we tested the following prosodic parameters in the realization of string-identical ISQs and RQs in German: nuclear pitch accent type, boundary tone, voice quality in the major constituents (verb, subject, and object noun in polar questions; *wh*-word, verb, and object noun in *wh*-questions), constituent durations (here operationalized as speech rate), and voice quality. In this paper, we focus on the usefulness of these parameters for the automatic classification of illocution type (RQ vs. ISQ). Given previous claims (often inaccurate) about the meaning of boundary tones in previous relevant literature [8, 16], our results are highly relevant to semantic modeling, as well as to the extraction of the functions of interrogatives in human-computer interaction.

[1] These particles also have a lexical meaning (*schon:* 'already', *auch*: 'also, too'). The lexical meaning does not trigger RQs.

2. Production data

2.1. Methods

2.1.1. Materials

We constructed 11 *wh*-interrogatives that fitted both a rhetorical and an information-seeking reading (e.g., *Who likes celery?*). To this end, we used predications that – out of context – may be true for some people and false for others (e.g., 'liking celery'). From these *wh*-interrogatives, we derived polar questions by replacing the *wh*-word by the indefinite pronominal subject *anyone* and adapted the syntactic structure to verb-first (V1). The polar questions thus contained an open element, similar to the *wh*-pronouns in *wh*-questions. In sum, we had 22 pairs of matched *wh*- and polar questions, henceforth referred to as interrogative pairs. Within the pairs, only the syntactic structure (*wh*-pronoun + verb vs. verb + subject) varied between question types, but the proposition expressed by the sentence radical was the same. Within RQs, the set of propositions denoted by the *wh*-interrogative and the set of propositions denoted by the polar interrogative with the indefinite subject are roughly the same.

For each interrogative pair, we constructed two contexts, one triggering an information-seeking interpretation of the interrogative and one triggering a rhetorical one. An example of the resulting quadruple is given in Table 1. To control for information structure and specifically to avoid effects of information structure on nuclear accent position and type, each context introduced the predication expressed in the sentence radical (e.g., *liking celery* in Table 1), rendering the referents of the constituents in the verb phrase discourse-given (see [18] for more details).

Table 1. *Contextual settings for polar and* wh*-questions in both illocution types (ISQ, RQ); contexts and target interrogatives are translated from German.*

ISQ	RQ
polar question	
You cooked a dish with celery. You would like to know whether your guests like this vegetable and will eat it or not. You say to your guests:	In the canteen they have casserole with celery on the menu. However, you know that nobody likes this disgusting vegetable. You say to your friends:
Mag denn jemand Sellerie? 'Does anyone like celery?'	
***wh*-question**	
You cooked a dish with celery. You would like to know which of your guests likes this vegetable and would like some of it. You say to your guests:	In the canteen they have casserole with celery on the menu. However, you know that nobody likes this disgusting vegetable. You say to your friends:
Wer mag denn Sellerie? 'Who likes celery?'	

The rhetorical contexts for a given interrogative pair (polar, *wh*) were identical. They all contained a sentence stating that it is generally known (or that the speaker knows) that nobody agrees with a certain proposition (e.g., *you know that nobody likes celery*). The information-seeking contexts differed from the rhetorical contexts in that they stated that the speaker was looking for some piece of information. The information-seeking contexts were largely identical for the two question types and differed only in whether uncertainty was expressed about the polarity (in polar questions; e.g., *whether or not your guests like it*) or about the subject (in *wh*-questions; e.g., *who likes it*). Each target interrogative ended in a mostly sonorous sentence-final object noun, consisting of two to four syllables with lexical stress on the penultimate or antepenultimate syllable. All target interrogatives contained the modal particle *denn,* which frequently occurs in both question types in German [19]. The use of *denn* facilitated the creation of natural target sentences in both conditions without biasing the interpretation of the utterance towards a rhetorical or information-seeking reading [20]. Hence, the illocution of the target was determined only by the contextual information.

As fillers, we used six questions with structural (PP-attachment) ambiguities, each of which occurred in two contexts. In addition, we constructed 22 exclamatives with V1 word order, i.e., the same word order as in polar questions.

2.1.2. Procedure

Two basic experimental lists were constructed. Each list contained the polar question for half of the question-pairs and the *wh*-question for the other half. Illocution type was manipulated within-subjects. That is, each participant produced both the rhetorical and the information-seeking version of each target interrogative, but only one question type of each interrogative pair. The 34 filler items were added to each list. The experimental lists were randomized anew for each participant with the constraint that two readings of a target interrogative were separated by at least four other trials. Each experiment started with four familiarization trials, followed by a short break in which participants were allowed to ask questions if anything was unclear. The experiment was controlled using the experimental software *Presentation* (Neurobehavioral-Systems, 2000). Each trial started with the visual display of the context, which the participant had to read silently, followed – upon button press – by the target interrogative on the next screen. The target sentence had to be produced aloud. Participants were asked to produce the questions in such a way that they were suitable in the given context. The experiment was self-paced. The recording started simultaneously with the appearance of the interrogative on the screen. After the production of the target, participants pressed a button to proceed to the next trial. The recording of the previous target was stopped at that point. Participants were allowed to repeat the question in case of mispronunciation or other mistakes (participants only rarely used this option, < 0.5% of the cases). No feedback was provided during the actual experiment. The experiment lasted about 25 to 30 minutes. Productions were recorded using a headset-microphone (Shure SM10A) and digitized directly onto a PC (44.1 kHz, 16Bit, stereo).

2.1.3. Participants

Twelve monolingual speakers of German (average age=21.7 years, SD=2.3; 10 female, 2 male) participated for a small payment. They were students at the University of Konstanz and unaware of the purpose of the study. The participants were randomly assigned to one of the two lists (6 in each list). None of them reported any speaking or hearing disorders.

2.1.4. Data Treatment

In total, we collected 528 target interrogatives (44 contexts x 12 participants), of which 26 realizations (4.9%) had to be excluded due to mispronunciation (N = 14), laughter (N = 2),

technical errors (N = 2) or audible pauses between the syntactic constituents (N = 8). In case of multiple recordings, the second recording was analyzed. The final data set consisted of 259 polar (RQ: 124, ISQ: 125) and 253 (RQ: 127, ISQ: 126) *wh*-questions.

The files were annotated at the word level using standard segmentation criteria [21] in the software package Praat [22]. Voice quality was classified as modal, breathy or glottalized in the initial word (verb in polar questions, *wh*-word in *wh*-questions), the second constituent (subject in polar questions, verb in *wh*-questions), and the final object noun. A perceptual classification was deemed more robust than acoustic measures, given variation in the materials regarding the quality of the stressed vowel and the word-prosodic structure of the words. For intonational analysis, pitch accents and edge tones were annotated according to the GToBI guidelines [23, 24]. The annotations were done by three trained annotators with substantial interrater reliability (kappa>0.71, [25]).

The continuous variables were analyzed with linear mixed effect regression models with illocution-type (ISQ vs. RQ) as fixed factor and participants and items as crossed random factors (adjustment of intercepts). Random slopes were added if this improved the fit of the models. P-values were calculated using the Satterthwaite approximation of degrees-of-freedom. Categorical variables were coded as 0 or 1 and analyzed using logistic mixed models. To avoid Type I errors, p-values were adjusted by means of the Benjamini-Hochberg correction [26].

2.2. Results

The production data resulted in the following differences between illocution types [18]. Phonologically, polar RQs typically ended in H-% (67%) and polar ISQs in H-^H% (88%). *Wh*-RQs almost exclusively ended in a low edge tone L-% (94%) whereas *wh*-ISQs exhibited more tonal variation (L-%: 44%, L-H%: 28%, H-^H%: 25%). Irrespective of question type, RQs were mainly realized with an L*+H nuclear accent (polar: 57%; *wh*: 57%), while polar ISQs were mostly realized with L* (81%) and *wh*-ISQs with L+H* (47%). Phonetically, irrespective of question type, RQs were realized more often with breathier voice quality than ISQs in the first word (37% vs. 7%, p<0.0005). Furthermore, RQs were on average 190ms longer than ISQs (p<0.0005), a lengthening of 15% relative to ISQs. The object noun was over-proportionally lengthened.

We trained separate classification and regression trees for the two question types (using the package rpart in R [27]). To this end, we excluded parameters that occurred less than 5 times in the illocution type where it was most frequent. The initial model included all phonological and phonetic properties. Instead of absolute object duration, which is dependent on the lexical material, we included speech rate (number of intended phones per second). Acoustic measures of voice quality (harmonics-to-noise ratio and H1*-A3*) were also included [28, 29]. To avoid overfitting, the resulting tree was pruned using the complexity parameter that resulted in the lowest cross-validation error [30]. To test the generalizability of the tree, we used a 10-fold cross-validation procedure (splitting the data in 10 random sets, training the tree on 8 sets and testing it on 2 sets). Accuracy was calculated on the 20% unseen data.

For polar questions, three data points with !H-% were excluded. Utterances ending in H-% were classified as RQ with only one exception (83 of 84 items were classified correctly). The other edge tones were mostly classified as ISQ (117 of 157). For the unseen data, classification accuracy was 87.5%.

For *wh*-questions, we removed eight data points (with the rare accent !H* and the rare boundary tones H-% and !H-%). The classification results showed that the initial split was caused by accent type, see Figure 1. Further factors were speech rate (duration of the final object) and the final boundary tone. An unseen test set (20% of the data) was classified correctly in 85.4% of the cases.

3. Discussion and Conclusion

The results of the production study showed that illocution type affected both intonational information (nuclear pitch accent types and boundary tones) and phonetic parameters (object duration and voice quality) in polar questions and *wh*-questions. For automatic classification, only the final boundary tone was used for polar questions. Utterances ending in H-% were classified as RQs in 99% of the cases; all other boundary tones as ISQs. RQs were classified more accurately than ISQs, which either suggests a very specific prosodic realization of RQs and more variability in ISQs, or a classification bias towards RQs, or both. Possibly, ISQs allow for the coding of other facets, such as politeness, emotional attitude, etc. Semantically, it is sufficient to use the boundary tone to model RQs, in line with previous semantic approaches, but against those approaches, it is a high plateau rather than a falling contour that specifies RQs. Similar intonational findings have been reported for English [31].

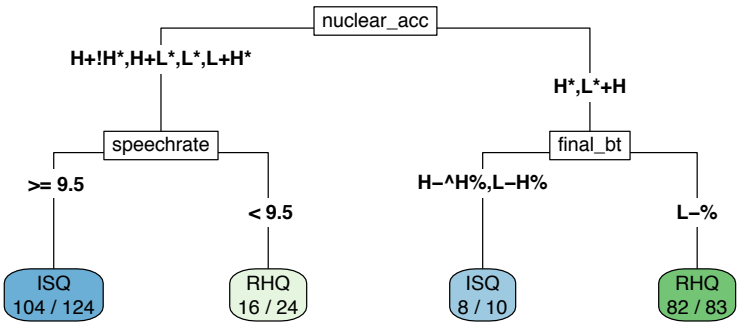


Figure 1. *Result of a pruned classification and regression tree for* wh*-questions. The values below the classified labels indicate the probability of the fitted class.*

For *wh*-questions, the initial split was caused by accent type, i.e., an intonational parameter. Further splits were due to accent type and speech rate. This suggests an interaction of phonological and phonetic cues for the classification of *wh*-questions. The interplay between phonetic and phonological parameters shows that *wh*-questions cannot be modeled based on intonational parameters alone [against 8, 16]. Recent perception studies point in the same direction. [32] tested the role of pitch accent type and voice quality in German *wh*-questions (L*+H L-% vs. H+!H* L-%, produced with breathy vs. modal voice quality). The L*+H L-% contour typically resulted in RQ judgments (with breathy voice: 93%, with modal voice over 61%), while H+!H* L-% resulted in mostly ISQ responses (modal voice: 92%, breathy voice: 72%). Hence, pitch accent type and voice quality are additive cues. In future work, we plan to use the current findings from automatic classification to derive further hypotheses for perception. In particular, the classification results suggest that speech rate may be a useful discriminator for certain accent types in *wh*-questions. Furthermore, we plan to test the classifier on non-experimental data. We will also include data from other languages to probe the language-specificity of these parameters.

4. Acknowledgements

This project was funded by the German Research Foundation (DFG) with grants to Braun (BR 3428/4-1) and Dehé (DE 876/3-1) as part of research unit "Questions at the Interface".

5. References

[1] Caponigro, I. and Sprouse, J., "Rhetorical questions as questions," *Proceedings of Sinn und Bedeutung 11*, Puig-Waldmüller, E., Ed., Barcelona: Universitat Pompeu Fabra, 2007, pp. 121-133.

[2] Hudson, R. A., "The meaning of questions," *Language,* vol. 51, pp. 1-31, 1975.

[3] Quirk, R., Greenbaum, S., Leech, G., and Svartvik, J., Eds., *A Comprehensive Grammar of the English Language*. New York, 1985.

[4] Wilson, D. and Sperber, D., "Mood and the analysis of non-declarative sentences," in *Human Agency: Language, Duty and Value*, Dancy, J., Moravcsik, J. M. E., and Taylor, C. C. W., Eds., Stanford, CA: Stanford University Press, 1988, pp. 77-101.

[5] Ilie, C., "The validity of rhetorical questions as arguments in the courtroom," in *Special Fields and Cases. Proceedings of the Third International Conference on Argumentation*, Amsterdam, 1995, pp. 73-88.

[6] Banuazizi, A. and Cresswell, C., "Is that a real question? Final rises, final falls, and discourse function in yes-no question intonation," *Proceedings of the 35th Annual Meeting of the Chicago Linguistics Society* Chicago, 1999, pp. 1-14.

[7] Biezma, M. and Rawlins, K., "Rhetorical questions: Severing asking from questioning," *Proceedings of SALT 27*, Burgdorf, D., Collard, J., Maspong, S., and Stefánsdóttir, B., Eds., 2017, pp. 302-322.

[8] Han, C.-H., "Interpreting interrogatives as rhetorical questions," *Lingua,* vol. 112, pp. 201-229, 2002.

[9] Ilie, C., "Rhetorical questions," in *The Pragmatics Encyclopedia*, Cummings, L., Ed., London, NY: Routledge, 2010, pp. 405-408.

[10] Sadock, J. M., *Toward a Linguistic Theory of Speech Acts*. New York, San Francisco, London: Academic Press, 1974.

[11] Sadock, J. M., "Queclaratives," in *Papers from the Seventh Regional Meeting, April 16-18, 1971*, Chicago Linguistics Society, 1971.

[12] Schaffer, D., "Can rhetorical questions function as retorts?: Is the Pope Catholic?," *Journal of Pragmatics,* vol. 37, pp. 433-460, 2005.

[13] Gutiérrez Rexach, J., "Rhetorical questions, relevance and scales," *Revista Alicantina de Estudios Ingleses,* vol. 11, pp. 139-155, 1998.

[14] Meibauer, J., *Rhetorische Fragen*. Tübingen: Nieymeyer, 1986.

[15] Wochner, D., Schlegel, J., Dehé, N., and Braun, B., "The prosodic marking of rhetorical questions in German," *Proceedings of the 16th Annual Conference of the International Speech Communication Association*, Dresden, Germany, 2015, pp. 987-991.

[16] Bartels, C., *The Intonation of English Statements and Questions. A Compositional Interpretation*. New York & London: Garland Publishing, 1999.

[17] Hedberg, N., Sosa, J. M., Gürgülü, E., and Mameni, M., "Prosody and pragmatics of wh-interrogatives," *Porceedings of the 2010 Meeting of the Canadian Linguistics Association*, 2010.

[18] Braun, B., Dehé, N., Neitsch, J., Wochner, D., and Zahner, K., "The prosody of rhetorical and information-seeking questions in German," submitted.

[19] Thurmair, M., *Zum Gebrauch der Modalpartikel 'denn' in Fragesätzen. Eine korpusbasierte Untersuchung*. Tübingen: Niemeyer, 1991.

[20] Thurmair, M., *Modalpartikeln und ihre Kombinationen*. Tübingen: Niemeyer, 1989.

[21] Turk, A., Satsuki, N., and Sugahara, M., "Acoustic segment durations in prosodic reserach: A practical guide," in *Methods in Empirical Prosody Research*, Sudhoff, S., Lenertová, D., Meyer, R., Pappert, S., Augurzky, P., Mleinek, I., *et al.*, Eds., Berlin, New York: De Gruyter, 2006, pp. 1-28.

[22] Boersma, P. and Weenink, D., "Praat: doing phonetics by computer," http://www.praat.org/, retrieved 11 May 2018, 2018.

[23] Grice, M. and Baumann, S., "Deutsche Intonation und GToBI," *Linguistische Berichte,* vol. 191, pp. 267-298, 2002.

[24] Grice, M., Baumann, S., and Benzmüller, R., "German intonation in autosegmental-metrical phonology," in *Prosodic Typology. The Phonology of Intonation and Phrasing*, Sun-Ah, J., Ed., Oxford: Oxford University Press, 2005, pp. 55-83.

[25] Landis, J. R. and Koch, G. G., "The measurement of observer agreement for categorical data," *Biometrics,* vol. 33, pp. 159-174, 1977.

[26] Benjamini, Y. and Hochberg, Y., "Controlling the false discovery rate: A practical and powerful approach to multiple testing," *Journal of the Royal Statistical Society,* vol. 57, pp. 289-300, 1995.

[27] Therneau, T. and Atkinson, B., "rpart: Recursive partioning and regression trees," https://CRAN.R-project.org/package=rpart, 2018.

[28] de Krom, G., "Some spectral correlates of pathological breathy and rough voice quality for differnet types of vowel fragments," *Journal of Speech and Hearing Research,* vol. 38, pp. 794-811, 1995.

[29] Keating, P. A. and Esposito, A., "Linguistic voice quality," *UCLA Working Papers in Phonetics U6,* vol. 105, pp. 85-91, 2007.

[30] Baayen, H. R., *Analyzing linguistic data. A practical introduction to statistics using R*. Cambridge: Cambridge University Press, 2008.

[31] Dehé, N. and Braun, B., "The prosody of rhetorical questions in English," submitted.

[32] Neitsch, J., Braun, B., and Dehé, N., "The role of prosody for the interpretation of rhetorical questions in German," *Proceedings of the 9th International Conference on Speech Prosody*, Poznan, Poland, 2018.

Production and perception of length contrast in lateral-final rimes

Tunde Szalay, Titia Benders, Felicity Cox, Michael Proctor

Department of Linguistics, Macquarie University, Sydney, Australia
ARC Centre of Excellence in Cognition and its Disorders
{tunde.szalay, titia.benders, felicity.cox, michael.proctor}@mq.edu.au

Abstract

Words containing /l/-final rimes challenge listeners as coda /l/ reduces certain vowel contrasts. Lateral-final rimes therefore allow us to gauge the link between individuals' word recognition and production. We tested whether participants producing a larger durational contrast between word pairs containing the rimes /iːl-ɪl, ʉːl-ʊl, æɔl-æl , əʉl-ɔl/ were better at recognising minimal pairs contrasting the aforementioned rimes. 46 Australian English speakers produced 24 /l/-final minimal pairs and identified the same minimal pairs spoken by two speakers. Participants producing a longer durational contrast took longer to respond and were only more accurate when the stimuli contained a bigger durational contrast.

Index Terms: durational vowel contrast, production, perception, lateral-final rimes, Australian English

1. Introduction

A growing body of experimental evidence shows that individuals' speech production and perception are linked [1–4]. Listeners who robustly produce a contrast are better able to perceive the same contrast than listeners with less robust contrast production [1–4]. For instance, listeners who more accurately differentiate voiced from voiceless stops also produce longer voice onset time [1]. Listeners who are better at discriminating the /s-ʃ/ contrast in perception maintain a more consistent tongue-tip contrast in production [2]. Listeners who are better at discriminating /ɑ-ʌ, u-ʊ/ produce greater spectral differentiation between members within these vowel pairs [3].

In perception, the phonological contrast between vowels is cued by several acoustic cues, i.e. formant values of vowel targets [5], vowel inherent formant change [6], and duration [5]. English, including Australian English, listeners rely more on spectral than durational contrast and use durational contrast only when spectral contrast is diminished or unavailable [5, 7]. That is, spectrally similar vowels are more likely to be confused [8] even when they differ in length [9].

Spectral contrast is weighted more heavily than durational contrast at an individual level [10]. Listeners from a speech community where spectral contrast is maintained between the vowels in PULL-POOL-POLE in the pre-/l/ context cannot discriminate these vowels in the speech of another speech community, where only durational contrast is maintained [10]. However, speakers who reduce spectral difference but maintain durational difference in production can utilize durational cues in perception even when spectral cues are not available [10]. This indicates that listeners rely on the same cues in perception which they produce and perceive as phonologically contrastive. In contrast, in the production and perception of voiced and voiceless consonants, listeners were found to weight voice onset time and $f0$ differently [11].

The aforementioned studies tested contrast perception on continua of manipulated stimuli, therefore little is known about if and how contrast production is associated with listeners' ability to cope with variation in unmanipulated speech. To further our understanding of the production-perception link, this study examined if and how contrast production is associated with word recognition and processing in Australian English (AusE) lateral-final rimes.

The AusE vowel inventory contrasts 18 stressed vowels, using both spectral and durational contrasts [14]. Some vowel pairs differentiated by duration exhibit smaller spectral differences (e.g. /ɐː-ɐ, iː-ɪ/, in *cart-cut*, *beat-bit*), others exhibit bigger spectral contrast (e.g. /ʉː-ʊ/, in *kook-cook*) [14]. There are diphthong-monophthong pairs in which the first or the second target of the diphthong coincides with a monophthong [14]. These inherent spectral similarities increase vowel confusion [9]. Coda /l/ further reduces the spectral contrast between /iː-ɪ, ʉː-ʊ, æɔ-æ, əʉ-ɔ/ (e.g. *feel-fill, fool-full, howl-Hal, dole-doll*); however contrastive duration may be maintained [15]. It is not clear if listeners can use durational differences in /l/-final rimes.

This study examined perception of duration contrast in CVl minimal pairs contrasting /iː-ɪ, ʉː-ʊ, æɔ-æ, əʉ-ɔ/ in the speech of two Source Speakers, one of whom maintains a more robust duration contrast than the other. The association between participants' production of the same duration contrast and their perception was tested in three hypotheses:

1. if AusE listeners rely on durational cues in /l/-final rimes, increased duration contrast in the stimuli would aid word recognition for all listeners regardless of their contrast production
2. if production and perception are linked, listeners producing a consistent length contrast would have an overall advantage in recognising /l/-final words that differ in the duration of the rime in the speech of both Source Speakers
3. if listeners rely more on cues that they themselves produce, then listeners who produce a more robust duration contrast would only perform better when the Source Speaker does so too.

2. Method[1]

2.1. Participants

Forty-six female [mean age = 21.5, range = 18 – 40] native speakers of AusE participated in the study. All participants were born in Australia to Australian-born parents. None of the participants reported any reading, hearing, or speaking disorders. Participants received course credit or $15 for participation.

[1] Data was collected as a part of a broader project.

2.2. Materials

The stimuli consisted of 32 CVC targets and 38 (C)V(C) fillers. 4 vowel pairs (/iː-ɪ, ʉː-ʊ, æɔ-æ, əʉ-ɔ/) were embedded in two sets of /l/-final and two sets of /d/-final minimal pairs to create 32 target words (See Table 1 for the /l/-final words). Here we analyse only production and perception data of /l/-final words.

Table 1: *Target words ending in /l/*

Vowel pair			
/iː-ɪ/	/ʉː-ʊ/	/æɔ-æ/	/əʉ-ɔ/
feel-fill, heel-hill	*fool-full, pool-pull*	*howl-Hal, vowel-Val*	*mole-moll, coal-Col*

To create the stimuli for the perception experiment, targets and fillers were read by two female native speakers (Source Speakers) of AusE upon orthographic random presentation on a computer monitor. Source Speaker 1 was 25, and Source Speaker 2 was 57 years old at the time of the recording. All stimuli were recorded with an AKG C535EB Condenser Microphone onto an iMac using Presonus Studio Live 16.2.4 AT Mixer in a sound treated studio. Stimuli were recorded at 44.1 KHz, amplitude-normalized, truncated to have 1 s silence before and after the word, and digitized as 16 bit WAV files.

Long:short rime duration ratios were calculated for the vowel-pairs /iː-ɪ, ʉː-ʊ, æɔ-æ, əʉ-ɔ/ from the experimental stimuli produced by the two Source Speakers (Table 2). Source Speaker 2 maintained a bigger long:short ratio, therefore maintained a bigger duration contrast for all vowel pairs except /æɔ-æ/.

Table 2: *Long:short rime duration ratios in the stimuli*

Informant	Vowel pair			
	/iː-ɪ/	/ʉː-u/	/æɔ-æ/	/əʉ-ɔ/
Source Speaker 1	1.27	1.3	1.23	1.23
Source Speaker 2	1.47	1.45	1.23	1.42

2.3. Procedure

The experiment consisted of a production task followed by a perception task, carried out in a one hour long session in a sound treated studio at Macquarie University, Sydney NSW. Participants were tested individually with the experimenter present.

Firstly, participants read orthographically presented words aloud. Words were pseudo-randomised, presented one by one three times in three blocks and recorded with an AKG C535EB Condenser Microphone onto an iMac using Presonus Studio Live 16.2.4 AT Mixer. The production task helped participants familiarise with the stimuli for the perception task.

Next, participants carried out the perception task, consisting of a practice phase and a test phase. In the practice phase, 10 single words were individually presented auditorily. Participants were asked to type the word that they heard quickly and accurately and received immediate feedback on what the correct responses were. In the test phase, participants were presented with individual words auditorily and were asked to type the words as they perceived them as quickly and accurately as possible. First, participants heard the words spoken by Source Speaker 2, repeated twice in two blocks, and then by Source Speaker 1, repeated twice in two blocks; blocks were separated by 30 s long forced break. Items within a block were pseudo-randomised so that no /l/-final words followed each other. Stimuli were presented with Expyriment [16] on an Asus X550JX laptop. Audio stimulus was presented via Sennheiser 380 Pro headphones at participants' preferred listening level. Participants' responses accuracy and response time (RT) of the first keypress were measured. After the word recognition task, participants were asked to fill out a self-evaluation questionnaire.

3. Data analysis

3.1. Production data

Recordings were segmented automatically [17]; rime durations were extracted automatically [18]. Rime duration is a measure combining vowel and coda /l/ length. Duration values 1.5 times above or below the interquartile range for a given vowel were hand-checked and corrected for measurement errors.

Mean rime duration was calculated by participant and vowel. The ratio of long:short vowels for each vowel pair and for each participant was calculated; increased ratio indicates an increased duration contrast.

3.2. Perception data

Responses to 46 (participants) x 64 (/l/-final tokens) = 2944 trials were collected. Responses were rated for accuracy. Responses were classified as Intended Answer, Phonetic Respelling, Typo, Minimal Pair Error, and Other Error. Responses were classified as Intended Answer if spelled as the target. Unambiguous but nonstandard phonetic spellings (e.g. *cole* for *coal*) were classified as Phonetic Respellings. Single letter deletions, additions, letter transpositions, and substitutions within one key distance of the target letter were classified as Typos [19], unless the result was an English lexical item. Confusion of members of minimal pairs (e.g. *fool* for *full*) was classified as Minimal Pair Error. Any other error (e.g. *cool* for *pool*, *howled* for *howl*) were classified as Other Error. For the purposes of the analysis of accuracy, Intended Answers, Phonetic Respellings and Typos were accepted as Correct; Minimal Pair Errors and Other Errors were classified as Incorrect.

RT of the first keypress was collected. RT within 210 ms [20] or above 5000 ms [21] of stimulus onset were excluded from analysis. Individual RT exceeding or less than mean+±2*sd for each participant were excluded from analysis [22]. 5.1% of responses were excluded according to these criteria, leaving 2,794 tokens for analysis.

4. Results

4.1. Individual variation in production and perception

Participants produced /l/-final rimes with a mean long:short ratio of 1.34 and a range of 0.99-1.38.[2] Participants consistently produced a decreasing durational contrast from /iː-ɪ/ to /ʉː-ʊ/ to /æɔ-æ/ to /əʉ-ɔ/. In the perception data, participants were consistent across the vowel pairs.

4.2. Production-perception link

To measure the association between accuracy, RT, and duration ratio, we constructed two Generalised Linear Mixed-effect models [23] with the dependent variables Accuracy and RT. The independent variables were Participant Duration Ratio (long:short, scaled), Vowel Pair (contrast coded and each compared against the grand mean), Source Speaker (contrast coded), and Lexical Frequency (from [24], log-normalised); Participant and Block were random intercepts. All two-way interactions

[2]Mean long:short vowel ratio was 1.64 in /d/-final rimes, as in [14].

between Duration Ratio, Vowel Pair, and Source Speaker were included in the model, but three-way interactions were not; Lexical Frequency did not interact with the other independent variables. Effects on accuracy were tested using the binomial family and effects on RT with the gaussian family with log-normal link, as raw RT followed a log-normal distribution.

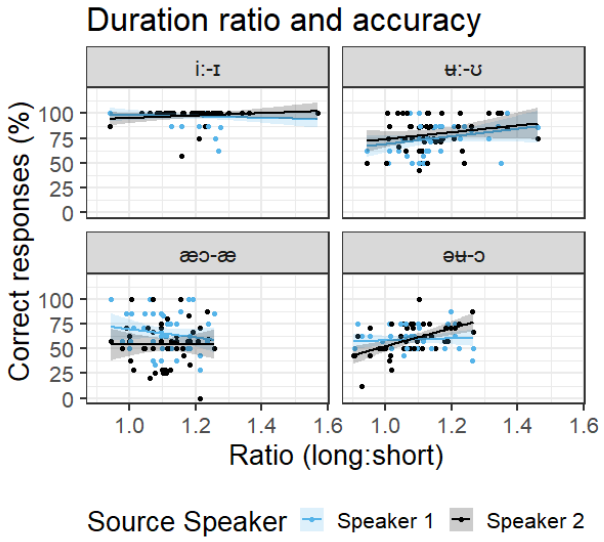


Figure 1: *Correlation of participants' duration ratio (x-axis) and recognition accuracy (y-axis) by Source Speaker (blue: Speaker 1, black: Speaker 2) and Vowel Pair (panels). Top: /iː-ɪ/ and /ʉː-ʊ/ contrast. Bottom: /æɔ-æ/ and /əʉ-ɔ/ contrast.*

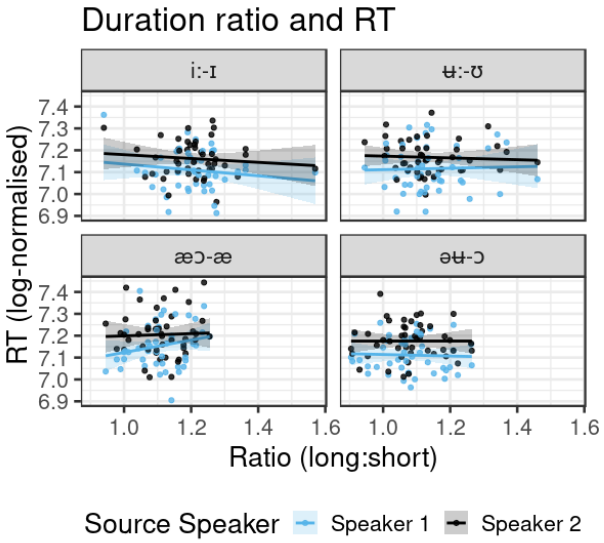


Figure 2: *Correlation of participants' duration ratio (x-axis) and perceptual RT (y-axis) by Source Speaker (blue: Speaker 1, black: Speaker 2) and Vowel Pair (panels). Top: /iː-ɪ/ and /ʉː-ʊ/ contrast. Bottom: /æɔ-æ/ and /əʉ-ɔ/ contrast.*

Participant Duration Ratio did not affect Accuracy significantly, but participants with larger Participant Duration Ratio had significantly slower RT (β=0.02, F(1, 4097)=9.53, p<0.001). Source Speaker did not affect Accuracy significantly, but participants responded more slowly to words produced by Source Speaker 2 (β=0.03, F(1, 4097)=0.0002, p=0.01). Participant Duration Ratio showed a significant positive interaction with Source Speaker 2 on accuracy (β=0.13, F(1, 5572)=9.74, p=0.002): participants with a larger long:short ratio recognised words more accurately when produced by Source Speaker 2, who produced larger duration contrast. Participant Duration Ratio and Source Speaker did not show significant interaction on RT.

Vowel Pair effects showed that /iː-ɪ/ was disambiguated more accurately (β=1.43, F(3, 5572)=105.95, p<0.001) and more quickly (β=-0.64, F(3, 4097)=99.11, p<0.001) than other Vowel Pairs. /ʉː-ʊ/ was disambiguated less accurately (β=-0.92, F(3, 5572)=105.92, p<0.001) but more quickly (β=-0.05, F(3, 4097)=99.11, p<0.001) than other Vowel Pairs. /ʉː-ʊ/ and Source Speaker 2 showed a negative interaction on RT (β=-0.02, F(3, 4097)=8.25, p<0.001): the RT difference between responses to Source Speaker 1 and 2 was smaller for /ʉː-ʊ/ than for other Vowel Pairs. /æɔ-æ/ was disambiguated less accurately (β=-0.18, F(3, 5572)=105.92, p=0.048) and more slowly (β=0.11, F(3, 4097)=99.11, p<0.001) than other pairs with 59% response accuracy and log-normalised 7.23 ms RT, in contrast with the overall response accuracy of 73% and log-normalised RT of 7.18 ms. Interactions between Participant Duration Contrast and Vowel Pair /æɔ-ɔ/ showed that participants with larger long:short ratio disambiguated /æɔ-ɔ/ less accurately (β=-0.22, F(3, 5572)=3.08, p=0.012) and more slowly (β=0.1, F(3, 4097)=1.69, p=0.04). Interaction between Source Speaker 2 and Vowel Pair /æɔ-æ/ showed that /æɔ-æ/ was disambiguated less accurately when produced by Source Speaker 2 (β=-0.25, F(3, 5572)=7.23, p=0.001).

Increased Lexical Frequency lead to increased accuracy (β=0.52, F(1, 5572)=256.3, p<0.001) and to increased RT (β=0.02, F(1, 4097)=63.23, p<0.001). Increase in RT with the increase in Lexical Frequency was probably due to the fact that there were more high frequency words among the targets with long acoustic duration.

4.3. Summary of findings

1. Contra to hypothesis 1, increased durational contrast in the speech of Source Speaker 2 did not assist word recognition, suggesting that not all listeners rely on durational cues.
2. Contra to hypothesis 2, participants who produced an increased durational contrast were not overall better at word recognition but they were overall slower.
3. In accordance with hypothesis 3, participants producing a larger duration contrast were more accurate on the contrast produced by Source Speaker 2, who, like them, maintained a larger durational contrast.

5. Discussion

Accuracy data showed that increased duration contrast in the stimuli aided word recognition only when participants also produced a more robust durational contrast. This indicates that perception is aided by cues that speakers themselves produce, but speaker-listeners without a robust durational cue production could not gain perceptual benefits. We found no evidence for overall better perception by participants with more robust duration contrast, contrary to [2, 3]. These discrepancies may be attributed to the differing methods, as we used an open-ended word recognition task, not contrast discrimination.

RT data showed that participants' increased rime duration contrast was associated with overall longer RT, indicating that these participants might consistently monitor for durational contrast. Durational contrast might take longer to process than spectral cues, as spectral cues may be available earlier in the

vowel, whereas the whole rime needs to be processed for the perception of durational cues [25, 26, c.f. 27]. The overall increase in RT with the increase in durational contrast in production indicates that speaker-listeners who rely on durational contrast in perception always monitor for it. However, the fact that these speaker-listeners are not overall more accurate indicates that they cannot always find durational contrast.

All participants responded more slowly to Source Speaker 2, despite Source Speaker 2 producing overall shorter target words than Source Speaker 1. The reason might lie in the potentially different spectral quality of the Source Speakers' vowels, in Source Speaker 2 always being presented first, or in the fact that Source Speaker 1 was closer in age to the participants.

Words contrasting the four vowel pairs were recognised differently and showed complex interactions with participants' production. Words contrasting /iː-ɪ/ were recognised more efficiently, potentially due to the F2 differences between /iː/ and /ɪ/ at vowel onset in the stimuli. Minimal pairs contrasting /æɔ-æ/ were poorly recognised, probably because neither of the Source Speakers produced a robust durational contrast for this vowel pair. All participants performed less accurately on Source Speaker 2's production of the /æɔ-æ/ contrast. Moreover, participants with a bigger durational contrast performed *worse* on the overall recognition of the /æɔ-æ/ contrast. That is, participants with bigger durational contrast did not perform better on Source Speaker 2, contrary to their performance with other vowel contrasts, as they may have been looking for a durational contrast that was not present. Patterns of minimal pair recognition contrasting /æɔ-æ/ are consistent with hypothesis 3, in which listeners' perception is aided by cues that they themselves produce.

These findings suggest that listeners can only benefit from durational cues in vowel perception when they themselves produce it. Similarly, in [10]'s study listeners who could not use durational contrast were members of a different speech community and maintained spectral contrasts (and presumably a non-phonological durational contrast as well), whereas participants in our study were members of a single speech community. These results do not allow us to determine the cues that listeners without a durational contrast use to identify /l/-final words. Future work will analyse listeners' spectral contrast production and link it to their perception of /l/-final minimal pairs.

6. Conclusion

Slower discrimination of /l/-final rimes by individuals who produce larger durational contrast implies that these speaker-listeners may monitor for durational contrast. This makes word identification slower, but only leads to increased accuracy when the speaker produces a sufficient durational contrast too. This implies that robust durational contrast production may come at a price and with limited benefits in word recognition.

7. Acknowledgements

We thank the Phonetics Lab at Macquarie University. This research was supported in part by iMQRTP 2015144, ARC DE150100318, and MQSIS 9201501719 grants.

8. References

[1] Newman, R., "Using links between speech perception and speech production to evaluate different acoustic metrics: A preliminary report", J. Acoust. Soc. Am., 113(5):2850–2860, 2003.

[2] Perkell, J. S., Matthies, M. L., Tiede, M., Lane, H. Zandipour, M., Marrone, N., Stockmann, E. and Guenther, F. H, "The distinctness of speakers' /s/-ʃ/ contrast is related to their auditory discrimination and use of an articulatory saturation effect", J. Speech Hear. Res., 47(6):1259–1269, 2004.

[3] Perkell, J. S., Guenther, F. H, Lane, H., Matthies, M. L., Stockmann, E., Tiede, M. and Zandipour, M., "Cross-subject correlations between measures of vowel production and perception", J. Acoust. Soc. Am., 116(4):2338–2344, 2004.

[4] Zellou, G., "Individual differences in the production of nasal coarticulation and perceptual compensation", J. of Phonetics 61:13–29, 2017.

[5] Bennett, D. C., "Spectral form and duration as cues in the recognition of English and German vowels", Language & Speech 11(2):65–85, 1968.

[6] Nearey, T. M. and Assmann, P. F., "Modelling the role of inherent spectral change in vowel identification", J. Acoust. Soc. Am., 80(5):1297–1308, 1986.

[7] Liu, S., "The effect of vowel duration on native Mandarin listeners' perception of Australian English vowel contrasts in voiced and voiceless coda contexts", M.Res. thesis, Dept. of Linguistics, Macquarie Univ., Sydney, NSW. 2016.

[8] Neel, A. T. "Vowel space characteristics and vowel identification accuracy", J. Speech Hear. Res., 51(3):574–585, 2008.

[9] Szalay, T., Benders, T., Cox, F. and Proctor, M., "Disambiguation of Australian English vowels", in C. Carignan and M. D. Tyler [Eds] Proc. of 16th Speech Sci. and Technol. Conf., 73-76, 2016.

[10] Wade, L., "The role of duration in the perception of vowel merger" J. of Laboratory Phonology 8(1), 2017.

[11] Shultz, A. A., Francis, A. L. and Llanos, F., "Differential cue weighting in perception and production of consonant voicing", J. Acoust. Soc. Am., 132(2):EL95–EL101, 2012.

[14] Cox, F., "The acoustic characteristics of /hVd/ vowels in the speech of some Australian teenagers", Australian J. of Linguistics 26(2):147–179, 2006.

[15] Palethorpe, S. and Cox, F., "Vowel modification in pre-lateral environments", Int. Seminar on Speech Prod., Sydney, 2003.

[16] Krause, F. and Lindemann, O., "Expyriment: A Python library for cognitive and neuroscientific experiments", Behaviour Res. Methods, 46(2):416–428, 2014.

[17] Schiel F., "Automatic phonetic transcription of non-prompted speech", in J. J. Ohala, Y. Hasegawa, M. Ohala, D. Granville, and A. C. Bailey [Eds], Proc. of the ICPhS, 607-610, 1999.

[18] Boersma, P. and Weenink, D., Praat: doing phonetics by computer v6.0.25, Online: http://fon.hum.uva.nl/praat/, accessed in 2017.

[19] Luce, P. A. and Pisoni, D. B., "Recognizing spoken words: The neighborhood activation model", Ear & Hearing 19(1):1-36, 1998.

[20] Woods, D. L., Wyma, J. M., Yund, E. W., Herron, T. J. and Reed, B., "Factors influencing the latency of simple reaction time", Frontiers of Human Neuroscience 131(9):1-12 2015.

[21] Baayen, H. R. and Milin, P., "Analysing reaction times", Int. J. of Psychological Res. 3(2):12–28, 2010.

[22] Ratcliff, R., "Methods for dealing with reaction time outliers", Psychological Bulletin 114(3):510-532, 1993.

[23] Bates, D., Mächler, M., Bolker, B. and Walker, S., "Fitting linear mixed-effects models using lme4", J. of Statistical Software 67(1):1–48, 2015.

[24] Davies, M., "Corpus of global web-based English: 1.9 billion words from speakers in 20 countries", Online: https://corpus.byu.edu/glowbe/, accessed in 2017.

[25] McMurray, B., Clayards, M. A., Tanenhaus, M. K. and Aslin, R. N., "Tracking the time course of phonetic cue integration during spoken word recognition", Psychonomic Bulletin & Review, 15(6), 1064-1071 2008.

[26] Reinisch, E. and Sjerps, M. J., "The uptake of spectral and temporal cues in vowel perception is rapidly influenced by context", J. of Phonetics, 41(2), 101-116 2013.

[27] Tillman, G., Benders, T., Brown, S. D. and van Ravenzwaaij, D., "An evidence accumulation model of acoustic cue weighting in vowel perception.", J. of Phonetics, 61, 1-12 2017.

Effects of glottalisation on reaction time in identifying coda voicing

Joshua Penney[1], *Felicity Cox*[1], *Anita Szakay*[1]

[1]Centre for Language Sciences, Department of Linguistics, Macquarie University

joshua.penney@mq.edu.au; felicity.cox@mq.edu.au; anita.szakay@mq.edu.au

Abstract

Australian English speakers employ both vowel duration and glottalisation to cue coda voicelessness. However, it is unclear how they use these cues in speech perception. Two groups of Australian English listeners (older/younger, *n*=77) responded to stimuli in which vowel duration and glottalisation were co-varied. Analysis of listeners' reaction times to stimuli suggests that glottalisation facilitates faster perception of coda voicing when paired with a congruent cue, but results in slower reaction times when paired with a competing cue. Despite age-based differences in production (older speakers use glottalisation less), in perception older and younger listeners used glottalisation in similar ways.

Index Terms: glottalisation, vowel duration, coda voicing, Australian English, reaction time.

1. Introduction

In syllable onset position, one of the major cues to voicing of singleton oral stops is voice onset time (VOT): the lag between the release of the stop closure and the onset of phonation [1]. In coda position, where VOT cannot apply unless there is a following voiced segment, one of the major cues to voicing is the duration of the preceding vowel. Vowels preceding voiced stops are longer than those preceding voiceless stops [2, 3, 4]. Additional acoustic cues such as differences in coda closure duration, differences in fundamental frequency (F0) at vowel offset, presence of voicing during the stop closure, and the presence or absence of glottalisation [5, 6] may also signal the coda stop voicing contrast.

Glottalisation is used to signal coda stop voicelessness in multiple varieties of English, for example, in American English (AmE), British English, Scottish English, and Australian English (AusE) [7, 8, 9, 10]. In AusE, preceding vowel duration and glottalisation may occur in a trading relationship. In a speech production study, [10] found that younger AusE speakers utilised preceding vowel duration as a voicing cue less than older AusE speakers, however they employed glottalisation more than the older speakers. Complementarily, both older and younger speakers produced less glottalisation in high vowel contexts, which in both age groups exhibited the greatest coda voicing related durational differences.

While glottalisation is used in voiceless coda production, little work has been conducted on how listeners utilise this cue in perception. In a phoneme monitoring task, [11] found that AmE listeners' reaction time (RT) and accuracy in identifying /t/ were improved when glottalisation was present, though listeners were not presented with voiced coda stops in that experiment. In an eye-tracking task, [12] also found that AmE listeners were marginally faster in identifing /t/ when it was glottalised compared to when it was not glottalised, though the effect was not significant. Furthermore, when words with voiceless and voiced codas were examined together, they found that glottalisation provided no benefit in the recognition of words with voiceless codas. On the other hand, glottalisation had an inhibitory effect on the identification of words with voiced codas, with listeners taking longer to identify words with voiced codas when glottalisation was present. A recent study examining the effect of co-varying vowel duration and glottalisation on AusE listeners found that glottalisation had a facilitative effect on the perception of coda voicelessness, with listeners providing more voiceless responses when glottalisation was present compared to when it was absent, even when this was paired with extended preceding vowel duration (i.e. a competing cue) [13]. Although this finding suggests glottalisation may strengthen the perception of coda voicelessness, it is not known whether the addition of glottalisation facilitates faster coda stop identification by AusE listeners. Thus, in this paper, we examine AusE listeners' RT to stimuli in which vowel duration and glottalisation are co-varied in order to explore whether the presence of glottalisation facilitates more rapid perception of coda stop voicelessness.

As older AusE speakers have been shown to utilise glottalisation less in production than younger speakers [10], in this study we also explore whether age-based differences are apparent in relation to listeners' RT. Differences in RT between older and younger groups are of course to be expected, as previous studies have shown that RT increases with age [14, 15]. However, our interest lies in whether older listeners exhibit a similar or different pattern of overall results to younger listeners: [13] found that, despite age-based differences in production, the presence of glottalisation had an enhancement effect for both older and younger listeners, as shown by an increased proportion of voiceless responses to stimuli containing glottalisation. Here we are interested in whether the presence of glottalisation also results in faster identification by listeners, in addition to facilitating perception of coda voicelessness as shown in [13]. Based on the observations above, we predict:

- Listeners' reactions will be faster for voiceless responses (and conversely slower for voiced responses) when vowel duration is short, and slower for voiceless responses (and conversely faster for voiced responses) when vowel duration is long.
- If glottalisation can facilitate listeners' perception of coda voicelessness, we predict that listeners' RT will be faster when glottalisation is present and vowel duration is short (i.e. as an additional cue congruent with a voiceless coda) but slower when it is present and vowel duration is long (i.e. as a competing cue).
- Listeners may also respond faster overall when glottalisation is present, regardless of vowel duration, if they only attend to the portion of the vowel that is modally voiced, as this is necessarily shorter in glottalised vowels than in non-glottalised vowels.

- Older listeners will demonstrate overall slower RT than younger listeners due to their age, but will show a similar general pattern, as in [13].

2. Methods

In order to examine the hypotheses we recorded listeners' RT to a two-alternative forced choice word identification task. Each participant was randomly presented with 648 items in which vowel and coda durations as well as the presence of glottalisation were manipulated. The subset of data analysed here comprises responses where only vowel duration and glottalisation were manipulated.

2.1. Participants

A total of 77 participants took part. Participants were allocated to one of two groups: an older or a younger group. The older group consisted of 31 listeners aged 50 years and above (f: 22; m: 9; mean age: 61); the younger group consisted of 46 listeners aged between 18 and 36 years (f: 38; m: 8; mean age: 21). All participants were native speakers of AusE and were either born in Australia or moved to Australia at a young age. All completed primary and secondary education exclusively in Australia and reported normal hearing for their age**.**

2.2. Stimuli

A 25 year old, female AusE speaker provided the original productions for the stimuli used in this study. The speaker was recorded in a sound treated recording studio at Macquarie University, using an AKG C535 EB microphone through a Presonus StudioLive 16.2.4 AI mixer to an iMac at 44.1kHz sampling rate. The speaker produced the words *bead*, *bard*, *bid* and *bud* in phrase medial position. All of the items were produced with modal voice. To enable the creation of an additional set of glottalised stimuli, sustained productions of each of the four vowels with creaky voice were also produced.

In order to create ambiguity regarding the voicing of the coda stops, we removed the original coda stops. These were then replaced with a low intensity coda stop release burst that had been produced in an unstressed syllable, and which our pilot tests had shown to produce ambiguous (i.e. neither consistently voiced nor voiceless) responses. In addition, we removed other acoustic cues to coda voicing by replacing the stop closure periods with silence and truncating the vowels before the F1 formant transitions into the following consonant began. Finally, we manipulated intensity contours to ensure consistency across all of the items, and manipulated F0 so that each token had a falling pitch contour ranging from 265hz at vowel onset to 203hz at vowel offset.

We then created a continuum for each item in Praat [16], in which vowel duration was increased in nine equally spaced steps. Minimum and maximum vowel durations were based on those reported for young females in [10]. The first (i.e. shortest) step corresponded to two standard deviations less than the mean duration for that vowel in a voiceless coda context, and the ninth (i.e. longest) step corresponded to two standard deviations greater than the mean duration for that vowel in a voiced coda context. The minimum and maximum durations for each continuum in milliseconds were as follows: bead: 104/340; bard: 165/370; bid: 65/166; bud: 64/201. The coda closure duration in each continuum was held constant and was based on the mean reported for each vowel context in [10]. In order to create an additional glottalised set of vowel duration continua, a duplicate set of continua was produced. However, in this set the final portion of each vowel was replaced with glottalisation (taken from the sustained vowels produced with creaky voice). For the inherently long vowels the final 25% of the vowel was replaced with glottalisation, and for the inherently short vowels the final 35% was replaced with glottalisation, consistent with previous reports [10].

2.3. Task procedure

Participants were presented with audio stimuli through Sennheiser HD 380 pro headphones while orthographic representations of minimal pairs differing only in coda voicing were presented on an Apple Macbook notebook. The minimal pairs were *bead/beat*, *bard/bart*, *bid/bit*, and *bud/but*. Prior to the perception task, all participants took part in a production task, in which all of these words (as part of a larger set) were produced in randomised carrier sentences. For the presentation of each audio stimulus a fixation cross was first displayed for 600 milliseconds in the middle of the screen, followed by the display of the orthographic representations of one of the minimal pairs (two words). After 500 milliseconds participants were presented with a single word audio stimulus over the headphones. Participants selected which word they heard by responding with a key press. Following the participant's response, the sequence began anew for the next stimulus. RT was measured from the onset of the audio presentation of the vowel to the participant's key press response.

All of the items presented had the form /bVC/, with /V/ one of the vowels /iː, ɪ, ɐː, ɐ/, and /C/ an alveolar stop with ambiguous voicing. Listeners were randomly assigned to either the inherently long vowel /iː, ɐː/ or the inherently short vowel /ɪ, ɐ/ condition, with each vowel presented in a separate block. This was necessary as the task would be too long and taxing for participants if they were to engage with the full set of stimuli. For each step in each continuum, a token without glottalisation and a token containing glottalisation were randomly presented. Participants were presented with three repetitions of each item.

2.4. Data analysis

We obtained RT data for 108 responses per participant (9 steps x 2 conditions x 2 vowels x 3 repetitions), resulting in a total of 8316 responses. RT responses below 200 milliseconds from the vowel onset were discarded, as this is below the time required to respond to a stimulus [17]. Similarly, responses that were greater than two standard deviations from the mean (>5261ms) were not included in the analysis. After trimming a total of 8223 responses remained for analysis.

We fitted a linear mixed effects model using *lme4* [18] in R [19]. The logRT of each response was included as the dependent variable. The fixed factors included were continuum step (from shortest to longest vowel duration), age group (older/younger), condition (non-glottalised/glottalised), and listener response (voiced/voiceless). Note that as inherently long and inherently short vowels were included in the data, as were high and low vowels (which also exhibit inherent durational differences), it was not possible to examine RT differences between the different vowels. Therefore, our analysis examines RT across all of the vowels. We also included as a control variable the logRT to the preceding stimulus, as doing so may improve model fit and assist in better estimation of the other factors of interest [20]. We also included interaction terms for condition, age, and step, and for

condition, response, and step. Random intercepts were included for participant and item.

3. Results

Table 1. *Summary of significant effects in linear mixed effects model.*

	Est	SE	*t*	*p*
preceding RT	0.149	0.01	22.28	<.0001
condition	-0.254	0.05	-4.57	<.0001
response	-0.225	0.03	-8.21	<.0001
step	-0.013	0.00	-2.85	0.0044
age	-0.109	0.05	-2.21	0.0291
conditionXresponse	0.150	0.05	2.79	0.0053
conditionXstep	0.044	0.01	5.48	<.0001
responseXstep	0.044	0.01	8.43	<.0001
conditionXageXstep	-0.012	0.01	-1.98	0.0477
conditionXresponse Xstep	-0.035	0.01	-4.13	<.0001

Table 1 displays a summary of significant effects in the model. As can be seen, the fitted model showed significant effects for preceding response RT, condition, response, step and age, as well as significant two-way interactions between condition and response, condition and step, and response and step. We also found significant three-way interactions between condition, age, and step and between condition, response and step. The significant effect for the control variable preceding response RT demonstrates that slower responses to a preceding stimulus resulted in a slower following response. In what follows we focus on the significant higher order interactions.

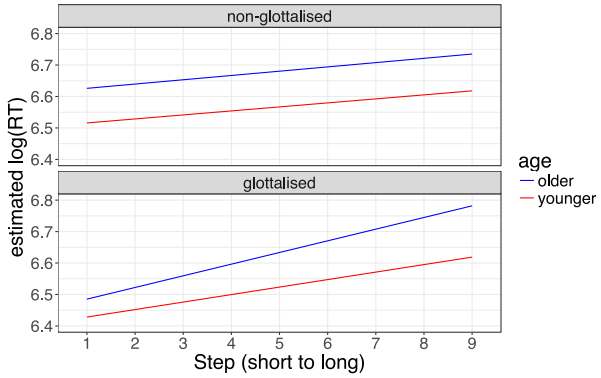


Figure 1: *Model estimates of logRT for older and younger listeners according to continuum step in non-glottalised (upper panel) and glottalised (lower panel) conditions.*

The three-way interaction between condition, age group, and continuum step suggests that there are differences in RT according to the continuum step that is presented between the older and younger listeners, and these differences also vary between the non-glottalised and glottalised conditions. Figure 1 shows the estimated logRT for each age group according to continuum step in both of the conditions. As can be seen, the older speakers have slower RT compared to the younger speakers at each step in both conditions, as is to be expected. However, the pattern of responses is similar between the older and younger groups. In the non-glottalised condition there is a gradual increase in RT as vowel duration increases, consistent with the fact that listeners may take longer to respond to longer tokens simply because the time between the stimulus presentation and the end of the vowel is longer. In contrast, in the glottalised condition we observe more extreme reactions, particularly at the lower end of the continuum: RT for the lower steps is faster than in the non-glottalised condition, whereas for the higher steps RT is the same as (younger), or slower than (older), in the non-glottalised condition. Post-hoc tests with Tukey correction show that, for the younger group, responses in the non-glottalised condition differ significantly from responses in the glottalised condition for steps 1-5 (p=0.008 and below), but not for steps 6-9. For the older group the result is similar, with significant differences between non-glottalised and glottalised for steps 1-5 (p=0.017 and below) and step 9 (p=0.046), but not for steps 6-8.

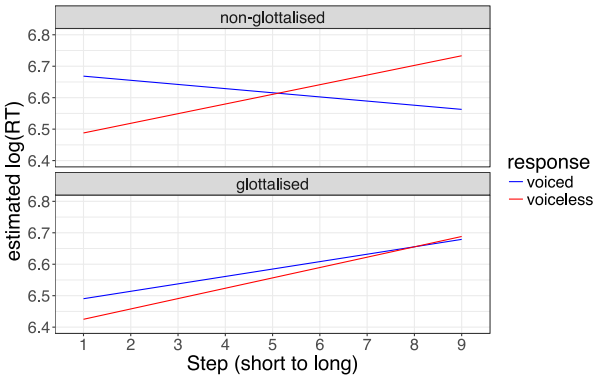


Figure 2: *Model estimates of logRT time for listener response according to continuum step in non-glottalised (upper panel) and glottalised (lower panel) conditions.*

The three-way interaction between condition, response, and continuum step suggests that listeners' RT varied according to which continuum step they were presented with and whether they selected a voiced or a voiceless coda, and that this differed between the non-glottalised and glottalised conditions. Figure 2 illustrates the estimated logRT for voiced and voiceless responses at each continuum step in each of the two conditions. As can be seen, in the non-glottalised condition listeners reacted faster with voiceless responses when the stimulus contained a shorter vowel (i.e. in the lower continuum steps), and their voiceless response RT increased as the vowel duration increased (i.e. in the higher continuum steps). For the voiced responses, on the other hand, the opposite pattern is observed: RT was slower when vowel duration was short, with a slight decrease as vowel duration increased. In the glottalised condition, however, the pattern is different: for both voiceless and voiced responses there is an increase in RT as vowel duration increases. Post-hoc tests with Tukey correction show that in the non-glottalised condition voiced and voiceless responses differ significantly from each another at each step (p=0.0027 and below) apart from steps 5 and 6. In the glottalised condition, however, there were no significant differences between voiced and voiceless responses.

4. Discussion

The results suggest that glottalisation does impact upon the time listeners require to respond to the stimuli. Overall, as expected, listeners responded more quickly when vowel duration was short than when vowel duration was long.

However, when glottalisation was included, listeners responded faster than when it was absent for the lower steps, but not for the higher steps. This suggests that glottalisation may facilitate faster perception of coda stops, but only when it is a congruent cue. That is, when glottalisation occurs with vowels of shorter duration (consistent with vowels preceding voiceless codas), i.e. the environment in which glottalisation is generally present, it improves listeners' RT. In contrast, when glottalisation is paired with longer vowels, it does not appear to benefit speed of identification. In addition, the fact that there was no decrease in RT for a voiceless stop percept in the glottalised compared to non-glottalised condition at the higher steps demonstrates that listeners were not faster to respond in the glottalised condition simply because the modally voiced portion of the vowel was shorter than in the non-glottalised condition. If this were the case, then we should expect to see faster reactions in the glottalised condition in the higher steps as well as in the lower steps, when compared with the non-glottalised condition (see Figure 1).

As predicted, older listeners were slower to respond than younger listeners, consistent with the literature on RT and age [14, 15]. Nevertheless, the effect of glottalisation decreasing RT for the lower but not the higher steps was observable in both age groups. This suggests that glottalisation affects speed of identification for older and younger listeners in the same way, similar to [13] where glottalisation promoted perception of voicelessness in both age groups, despite older speakers being less likely to utilise glottalisation in production.

When listeners' responses (i.e. selecting a voiced or voiceless coda) are considered, in the non-glottalised condition participants responded faster when selecting a voiceless response when vowel duration was short, and slower when vowel duration was long. Conversely, voiced responses were slower when vowel duration was short and faster when vowel duration was long (despite the fact that the stimulus was longer in the higher continuum steps and as such we might expect an increase in RT). This suggests that, as expected, listeners make use of the vowel duration cue, with congruent vowel duration reducing the time taken to respond, and incongruent vowel duration resulting in increased RT. The fact that there were voiced responses to the lower steps and voiceless responses to the higher steps suggests that (at least some) listeners were not basing their judgements solely on vowel duration, though it should be noted that these accounted for a only small proportion of responses (see [13] for more details).

When glottalisation was present, RT for voiceless responses was faster (than when glottalisation was not present) at the lower steps, supporting the claim that the addition of glottalisation improves listeners' perception of coda voicelessness, but only when vowel duration is not a competing cue. This result appears contrary to [12], where glottalisation was not found to improve listeners' reactions to words with voiceless codas, though it should be borne in mind that in [12] the coda stop voicing was not intentionally ambiguous as it was here.

For voiced responses to the higher steps with glottalisation we saw an increase in RT, consistent with listener conflict upon being presented with competing cues, and also consistent with results in [12], suggesting that listeners associate glottalisation with voiceless but not voiced codas. Curiously, however, the addition of glottalisation improved rather than delayed listener RT for voiced responses to the shorter steps, despite the fact there were two cues to voicelessness in the stimuli (i.e. shorter preceding vowel and glottalisation). This may be explained by the low number of voiced responses at the lower steps in the glottalised condition (6% voiced responses to steps 1-4 in glottalised condition vs 31% in non-glottalised condition). It is also possible that these responses were made in error.

Overall, we found that listeners' reactions times were improved by the addition of glottalisation, but only when this was paired with vowels of short duration. When vowel duration was extended, the addition of glottalisation did not result in faster reactions. This effect was visible for both older and younger listeners, suggesting that both age groups utilise the glottalisation cue to coda voicing similarly in perception.

5. References

[1] Lisker, L. and Abramson, A. S., "A cross language study of voicing in initial stops: Acoustical measurements", Word, 20:384–422, 1964.

[2] Fowler, C. A., "Vowel duration and closure duration in voiced and unvoiced stops: There are no contrast effects here", JPhon., 20:143–165, 1992.

[3] Klatt, D. H., "Linguistic uses of segmental duration in English: Acoustic and perceptual evidence", JASA, 59:1208–1221, 1976.

[4] Raphael, L. J., "Preceding vowel duration as a cue to the perception of the voicing characteristic of word-final consonants in American English", JASA, 51:1296–1303, 1972.

[5] Lisker, L., "'Voicing' in English: A catalogue of acoustic features signaling /b/ versus /p /in trochees", Lang. and Speech, 29, 3–11, 1986.

[6] Song, J. Y., Demuth, K. and Shattuck-Hufnagel, S., "The development of acoustic cues to coda contrasts in young children learning American English", JASA, 131:3036–3050, 2012.

[7] Roach, P. J., "Glottalization of English /p/, /t/, /k/ and /ʧ/ – a re-examination", JIPA, 3:10–21, 1973.

[8] Pierrehumbert, J., "Prosodic effects on glottal allophones", in *Vocal fold physiology: Voice quality control*, O. Fujimura and M. Hirano, Eds. San Diego: Singular, 1995, pp. 39–60.

[9] Gordeeva, O. B. and Scobbie, J. M., "A phonetically versatile contrast: Pulmonic and glottalic voicelessness in Scottish English obstruents and voice quality", JIPA, 43: 249–271, 2013.

[10] Penney, J., Cox, F., Miles, K. and Palethorpe, S., "Glottalisation as a cue to coda consonant voicing in Australian English", JPhon., 66:61–184, 2018.

[11] Garellek, M., "The benefits of vowel laryngealization on the perception of coda stops in English", UCLA Working Papers in Phonetics, 109:31–39, 2011.

[12] Chong, A. and Garellek, M., "Online perception of glottalized coda stops in American English," Laboratory Phonology, 9:1–24, 2018.

[13] Penney, J., Cox, F. and Szakay, A., "Weighting of coda voicing cues: Glottalisation and Vowel Duration", Proc. of INTERSPEECH, 1422–1426, 2018.

[14] Welford, A. T., "RT, speed of performance, and age", Annals NY Academy of Science, 515:1–17, 1988.

[15] Der, G. and Deary, I. J., "Age and sex differences in RT in adulthood: Results from the United Kingdom health and lifestyle survey", Psychology and Aging, 21(1): 62–73, 2006.

[16] Boersma, P. and Weenik, D., "Praat: Doing phonetics by computer [Computer program]", Version 5.4.09, available from http://www.praat.org/.

[17] Woods, D. L., Wyma, J. M., Yund, E. W., Herron, T. J. and Reed, B., "Factors influencing the latency of simple RT", Frontiers in Human Neuroscience, 9:1–12, 2015.

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

[19] R Core Team "R: A language and environment for statistical computing [Computer program]", Version 3.3.1, available at https://www.R- project.org/.

[20] Baayen, R. H. and Milin, P., "Analyzing RTs", International Journal of Psych. Research, 3:12–28, 2010.

Acoustic correlates of prominence in Nafsan

Rosey Billington[12], Janet Fletcher[12], Nick Thieberger[12], Ben Volchok[2]

[1]Centre of Excellence for the Dynamics of Language
[2]The University of Melbourne

rbil@unimelb.edu.au; janetf@unimelb.edu.au; thien@unimelb.edu.au; volchokb@unimelb.edu.au

Abstract

Though Oceanic languages are often described as preferring primary stress on penultimate syllables, many different patterns have been noted across and within language families, and may interact with segmental and phonotactic factors. This is exemplified across linguistically diverse Vanuatu. However, both impressionistic and instrumentally-based descriptions of prosodic patterns and their correlates are limited for languages of this region. This paper presents preliminary acoustic and durational results for Nafsan, an Oceanic language of Vanuatu, which suggest a preference for prominence at the right edge of words, with fundamental frequency as a primary correlate.

Index Terms: Oceanic, stress, duration, f0, intensity, F1, F2

1. Introduction

Nafsan (South Efate) is a Southern Oceanic language spoken by an estimated 6,000 people in three villages (Erakor, Eratap and Pango) on the island of Efate in Vanuatu. Though the phonology of Nafsan has been discussed in comparative and descriptive work [1] [2] [3], some challenges remain in understanding Nafsan segmental and prosodic patterns. Key questions are whether there is a vowel length distinction, why some vowels undergo deletion, what patterns of word- and phrase-level prominence are used, and how different parts of the sound system interact. This study is part of a wider project using instrumental phonetic approaches to address these questions, and presents selected results focusing on prominence patterns within words.

1.1. Segmental inventory of Nafsan

Nafsan has an inventory of 15 consonant phonemes, and has five contrastive vowel qualities, as is typical of Oceanic languages [3] [4]. Though the possibility of a length distinction has been mentioned in previous work [2], the status of length within the Nafsan vowel system was until recently unresolved. New phonetic data allowing for a targeted investigation of Nafsan vowels provides evidence for a monophthong inventory comprising /i, i:, e, e:, a, a:, o, o:, u, u:/; each of the five vowel qualities may occur either phonemically short or long, in various syllable types. Ongoing experimental work suggests that at least in CVC syllables, long vowels are close to twice as long as short vowels, and the duration difference between long and short vowels is approximately the same across all five vowel qualities [5].

1.2. Vowel deletion and phonotactic patterns in Nafsan

Nafsan phonotactic patterns are strikingly complex both compared to patterns for languages spoken further to the north in Vanuatu, and compared to the more typologically common preference for CV syllables among Oceanic languages [2] [4]. The language exhibits a range of heterorganic consonant clusters in syllable onsets, with various possible sonority profiles [3] [5]. There is some evidence, through comparisons with historical records and with cognates in closely-related languages, that these complex syllable onsets may have arisen through the deletion of selected medial vowels, but the status of and environments for vowel deletion have remained unclear [3]. Ongoing work indicates that vowel deletion is both a historical and productive process, and that at least for the productive process as used by contemporary speakers, short vowels are frequently deleted when they occur in the penultimate syllable of the word, though this appears to be mediated by lexical, grammatical, and speaker-specific factors [6]. These observations raise the question of whether vowel deletion in Nafsan, when it occurs, is pre-tonic, but given that prominence patterns in Nafsan remain under-described, this has not been clearly established.

1.3. Stress in Oceanic languages

It is often observed that many Oceanic languages display primary stress on penultimate syllables [4], but crosslinguistic examinations suggest that this tendency is not as widespread as previously thought [7]. Contemporary Oceanic languages exhibit a range of prominence patterns, including stress which is regularly penultimate, generally penultimate but final if the final syllable contains a coda and/or a long vowel or diphthong, final, initial, antepenultimate, lexically specified, or dependent on morphological factors [7]. All of these have been reported for at least some of the 130+ languages of Vanuatu, and varied prominence patterns are noted even across closely-related languages, such as Nafsan and neighbouring varieties [8, 9, 10]. For Nafsan, previously suggested patterns include initial and final stress, though the unclear status of vowel length has been a complicating factor [3, 11], and recent auditory impressionistic analyses suggest possible final prominence [5], but this has not yet been examined experimentally. Crosslinguistically, various acoustic and durational cues may correlate with lexical prominence [12], but for the languages of Vanuatu, impressions of stress correlates are only occasionally noted, and have not yet been supplemented by instrumental data. Perceived correlates include combinations of increased duration, pitch, and intensity/loudness for stressed vowels or syllables [13, 14, 10], but in some cases, for languages with a vowel length contrast such as Anejom̃, duration may not be a salient cue [15, 16].

2. Research aims

Given that recent work establishes that vowel length is distinctive in Nafsan, and that vowel deletion patterns and auditory impressions raise the possibility of word-final prominence, two key questions emerge. Are there phonetic differences in the realisation of vowels in final syllables, on the basis of duration, intensity, fundamental frequency, and formant frequencies, which suggest they are more prominent than preceding vowels? Relatedly, do long and short vowels show similar phonetic characteristics in the same word position or are they treated differently?

3. Method

3.1. Participants

The participants in this study were four adult speakers of Nafsan from Erakor village in Efate, Vanuatu: three men (GK, LE, MJ) and one woman (MK). All identify Nafsan as their first language, and the language they use at home. In addition, all speak Bislama, the English-lexified creole which is a lingua franca in Vanuatu, have knowledge of either English and a little French or French and a little English, and also have some knowledge of language varieties spoken in other parts of Vanuatu.

3.2. Materials and procedures

A set of two-syllable and three-syllable word forms was compiled as stimuli, drawing on existing databases and corpora for Nafsan. The words were selected to comprise only CV(C) structures (no complex onsets), and to have long and short vowels in different word positions, to allow investigation of whether these different phonotactic structures influence potential prominence patterns. The majority of the vowels were open /a, a:/. Results presented here pertain to two-syllable words only, with the following structures: CV.CVC, CV.CVVC, CVV.CVC, and CVV.CVVC. Examples include /rakat/ 'bite (DUAL)', /rapa:k/ 'delouse (DUAL)', /ka:kas/ 'be sweet', and /ta:k͡pa:r/ 'sin'. The duration values of vowels in initial and final syllables are not directly comparable in these structures, given that initial syllables are open and the final syllables are closed, but it is the characteristics of non-final CV syllables which are of particular interest, given that this is an environment in which vowel deletion occurs (though deletion is not attested for vowels in the word forms included here). In addition, though medial CVC syllables are possible in Nafsan, they are less frequent, and word-final CV syllables are much less common than final CVC syllables.

Each word was recorded three times in a medial frame, following a spoken prompt; the frame was *komam util* _____ *sernrak* 'We say _____ usually'. Recordings were made in a sheltered area during fieldwork in Erakor, and have been archived with other recordings relating to the wider project on Nafsan phonetic and phonological patterns [17]. Data were recorded at an archival sampling rate of 96kHz and 24-bit depth, using a Zoom H6 audio recorder and a Countryman H6 headset microphone with a hypercardioid polar pattern, and downsampled to 44.1kHz 16-bit for analysis. Tokens were balanced across speakers but not vowel length category and word structure (Table 1), given the limitations of available lexical data and the frequency with which different structures occur. The final dataset contained 1,114 vowel tokens drawn from 557 utterances produced (containing 46 different Nafsan words as the target word).

Table 1: *Number of tokens in dataset.*

word shape	initial /V/	final /V/	initial /V:/	final /V:/
CV.CVC	229	229	-	-
CV.CVVC	125	-	-	125
CVV.CVC	-	128	128	-
CVV.CVVC	-	-	75	75
total	**354**	**357**	**203**	**200**

3.3. Data processing and analysis

Utterances were transcribed orthographically in Praat [18], and orthographic transcriptions were converted to phonemic transcriptions in the Speech Assessment Methods Phonetic Alphabet (SAMPA). Using the TextGrid files and associated WAV files, automatic segmentation of the speech signal was performed via the web interface of the Munich Automatic Segmentation System (WebMAUS) [19], using the language-independent model for segment identification. Segment boundaries in the output TextGrid files were checked and manually corrected where necessary with reference to wideband spectrograms and corresponding waveforms in Praat. A hierarchical database was constructed using the EMU Speech Database Management System [20], including tiers for the phonemic segments, syllables, and words. The acoustic and durational characteristics of vowel tokens produced in the target words were queried and analysed using the `emuR` package in R [21, 22]. Measures of interest in this study are vowel duration (in ms), intensity at vowel midpoints relative to the midpoint of the (typically light) coda lateral in the preceding word (root mean square amplitude, in dB), fundamental frequency at vowel midpoints (in Hz), and first and second formant frequencies at vowel midpoints (in Hz). The data were tested with linear mixed-effects models using the `lme4` package [23] with random slopes and intercepts for speaker and word. Model validity was checked using a likelihood ratio test, and differences are reported based on Tukey's Honest Significant Difference post-hoc tests.

4. Results

4.1. Duration

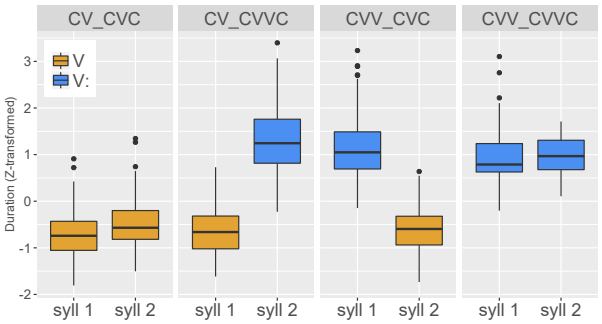


Figure 1: *Duration (Lobanov-normalised) of short and long vowels in initial and final syllables of disyllabic Nafsan words (by word structure).*

Duration values for vowels, shown in Figure 1, clearly illustrate the vowel length distinction across the different word structures in this data. The likelihood ratio test shows that there is a significant effect of context ($\chi^2(7)= 1304$, $p<0.001$), but post-hoc tests reveal that almost all significant differences are between vowels whose phonemic length differs. Final long vowels are an estimated 72 $\pm$ 2 ms longer than initial short vowels in CV.CVVC words ($p<0.001$), and initial long vowels are an estimated 61 $\pm$ 2 ms longer than final short vowels in CVV.CVC words ($p<0.001$). There are no significant differences between long vowels in any of their four contexts. For short vowels, there is a small but significant difference of 8 $\pm$ 2 ms between initial and final short vowels in CV.CVC words ($p<0.001$), but there are no other notable differences. Though larger differences may be observed for initial and final vowels of the same quantity in data with open rather than closed final syllables, the distributions shown here suggest that any duration differences correlating with word position are likely much smaller than those correlating with phonemic length. Vowel quality was not found to be a significant factor affecting duration when statistical models were compared (recalling that most tokens, 91%, were /a, a:/). Short vowels were on average 61ms in initial syllables and 66ms in final syllables, and long vowels 126ms in initial syllables and 134ms in final syllables.

4.2. Intensity

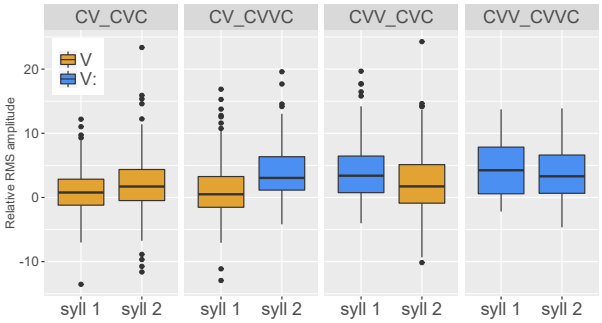


Figure 2: *Root mean square amplitude at midpoints of short and long vowels (relative to midpoints of sonorant coda in preceding word) in initial and final syllables of disyllabic Nafsan words (by word structure).*

Intensity patterns for vowels, based on the difference in root mean square (RMS) amplitude between the vowel midpoint and the midpoint of the sonorant coda in the preceding word, are shown in Figure 2. The effect of context is significant ($\chi^2(7)= 88$, $p<0.001$). However, post-hoc tests show that there are no significant differences between vowels of the same phonemic length in CV.CVC words ($p=0.14$) and CVV.CVVC words ($p=0.97$), nor between any vowels of the same phonemic length in other comparisons. For CV.CVVC words, there is a significantly larger increase in intensity for the final long vowels than the initial short vowels, by an estimated 2.9 ± 0.4 dB ($p<0.001$). For CVV.CVC words there are also significant differences but with higher values for the initial long vowel rather than the final vowel, of an estimated 1.7 ± 0.4 dB ($p<0.001$). As for duration, then, differences on the basis of RMS amplitude correlate more with vowel length than word position.

4.3. Fundamental frequency

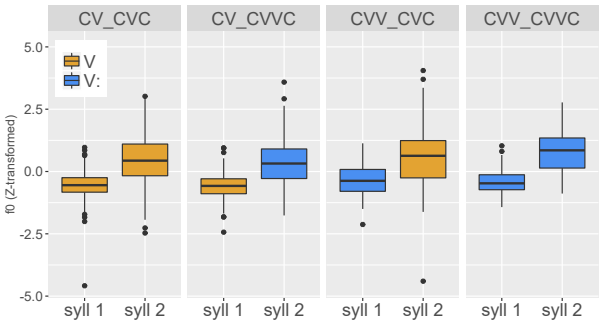


Figure 3: *Fundamental frequency (Lobanov-normalised) at midpoints of short and long vowels in initial and final syllables of disyllabic Nafsan words (by word structure).*

For fundamental frequency (f0), measured at the vowel midpoint, a subset of the data was used; 74 vowel tokens which returned zero values, due to breathiness or devoicing of the vowel, were excluded. Of these, 70% were from the same speaker, whose speech rate was noticeably faster than that of the other participants. 92% were short vowels; 45% of these were short vowels in initial syllables, and 49% were short vowels in final syllables. Results for the remaining 1,040 tokens are shown in Figure 3, and a consistent pattern can be seen across the four word structures. The effect of context is significant ($\chi^2(7)= 361$, $p<0.001$), and post-hoc tests confirm that there are significantly higher f0 values in final compared to initial syllables in each case. The differences are also of a similar magnitude; vowels in final syllables are an estimated 15 ± 1 Hz higher in CV.CVC words ($p<0.001$), 14 ± 2 Hz higher in CV.CVVC words ($p<0.001$), 14 ± 1 Hz higher in CVV.CVC words ($p<0.001$) and 19 ± 2 Hz higher in CVV.CVVC words ($p<0.001$). There are no significant f0 differences between vowels of the same phonemic length occurring in the same word position.

4.4. First and second formant frequency

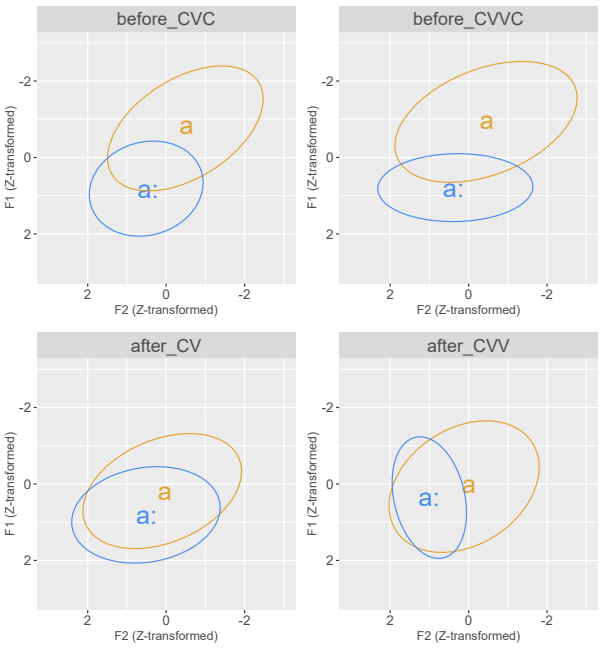


Figure 4: *First and second formant frequency (Lobanov-normalised) at midpoints of short and long vowels in initial and final syllables of disyllabic Nafsan words (by word position).*

A subset of the data containing only open vowels /a, aː/, which as noted comprised the majority of the data (1,012 tokens), was used to examine any differences in first formant frequency (F1) and second formant frequency (F2). Results for F1 and F2 measured at the vowel midpoint are shown in Figure 4. Likelihood ratio tests show that the effect of context is significant for both F1 ($\chi^2(7)= 587$, $p<0.001$) and F2 ($\chi^2(7)= 217$, $p<0.001$). The results of post-hoc tests include significant differences between short and long vowels in initial syllables; preceding CVC syllables, short /a/ has F1 values an estimated 189 ± 12 Hz lower than long /aː/ ($p<0.001$), and F2 values an estimated 144 ± 31 Hz lower ($p<0.001$). Preceding CVVC syllables, short /a/ similarly has F1 values an estimated 196 ± 21 Hz lower than long /aː/ ($p<0.001$), but F2 differences are not significant ($p=0.16$). In final syllables, the formant characteristics of /a/ and /aː/ differ less. Following CV syllables, F1 values for /a/ are an estimated 76 ± 13 Hz lower than for /aː/ ($p<0.001$); this is smaller than the quite substantial F1 difference in initial syllables, and there are no significant F2 differences between /a/ and /aː/ in this context ($p=0.6$). Following CVV syllables, there are no significant differences between short and long vowels on the basis of F1 ($p=1.00$) or F2 ($p=0.5$).

5. Discussion and conclusions

These findings provide compelling evidence that disyllabic Nafsan words are more prominent at the right edge. The consistent pattern of higher f0 values in final compared to initial syllables, regardless of the phonemic length of the vowel, suggests that f0 likely plays an important role in prominence marking. Results for F1 and F2 at midpoints of vowels in different contexts offer supporting evidence for right-edge prominence; in final syllables, the spectral differences between /a/ and /aː/ are minimal, but in initial syllables, F1 values are substantially lower for short /a/, and F2 values are somewhat lower, indicating centralisation and some degree of retraction for these short vowels. This is of particular interest given impressions that where productive vowel deletion occurs, it is generally of penultimate short vowels in CV syllables [2, 3, 6]. Penultimate short vowels in this dataset are not deleted, but still show some reduction.

Duration and intensity are likely only minor cues to prominence; the only difference between vowels of the same phonemic length was for the duration of initial and final short vowels in CV.CVC words, and the size of the estimated difference was very small in this case. Instead, duration and intensity appear to be robust correlates of vowel length and bolster evidence for the status of quantity distinctions in Nafsan [5]. Vowels identified as being phonemically long show reliably and substantially higher duration values than short vowels in different contexts, as well as higher intensity values as indicated by RMS amplitude at vowel midpoints. While there may be larger duration differences correlating with word position in data where both the initial and final syllables are either open or closed, duration differences correlating with vowel length are likely to be preserved regardless of syllable prominence. Similar observations have been made based on exploratory duration measures for vowels in the closely-related language Lelepa [10].

The results presented here accord with other impressions of final prominence in Nafsan [11, 5], and in pointing towards the importance of f0 as a major correlate, also suggest the possibility that prominence marking in Nafsan is more like that of languages such as Japanese which have been described as having non-stress accent [24]. Given that other preliminary work on Nafsan indicates that high f0 targets at the right edge of a word may demarcate the right edge of an accentual phrase, it will also be worth considering whether the language has a prosodic system like that of Korean or French, with high tone targets at right edges relating to constituents that are not necessarily lexical. This is the subject of ongoing research. The present results for vowels produced in an utterance-medial frame controlling for word length and syllable structure will provide a useful point of reference in ongoing work, which includes comparisons of the phonetic characteristics of Nafsan words of different lengths and syllable structures in initial, medial and final contexts.

Oceanic languages are under-represented in prosodic research, and as overviews of proposed prominence patterns show, there is much that remains to be understood [7]. The languages of Vanuatu are especially well-suited to investigations of the different ways that segmental and prosodic phenomena interact; they show enormous diversity in their sound systems, and appear to have responded in different ways to various sound changes which can be traced back to Proto-Oceanic. These results show that a more detailed understanding of language-internal prosodic patterns may offer insights into the phonetic mechanisms underpinning historical changes, while also contributing to a better understanding of the typological profile of these languages.

6. Acknowledgements

Sincere thanks to all the Nafsan speakers who have participated in and facilitated this and earlier work, in particular Gray Kaltaβau, Lionel Emil, Michael Joseph and Marinette Kalpram for their involvement in this study, and Carol Lingkary and Yvanna Ataurua for their assistance during fieldwork in Erakor. This research was conducted with support from the ARC Centre of Excellence for the Dynamics of Language (Project ID: CE140100041).

7. References

[1] Clark, R. "The Efate dialects", Te Reo, 28:3–35, 1985.

[2] Lynch, J. "South Efate phonological history", Oceanic Linguistics, 39(2):320–338, 2000.

[3] Thieberger, N. A grammar of South Efate: An Oceanic language of Vanuatu. University of Hawaii Press, 2006.

[4] Lynch, J., Ross, M., and Crowley, T. The Oceanic languages. Curzon Press, 2002.

[5] Billington, R., Fletcher, J., and Thieberger, N. "Nafsan", submitted.

[6] Billington, R., Thieberger, N., and Fletcher, J. "Phonetic evidence for phonotactic change in Nafsan", submitted.

[7] Lynch, J. "Reconstructing Proto-Oceanic stress", Oceanic Linguistics, 39(1):53–82, 2000.

[8] Schütz, A. Nguna grammar. University of Hawaii Press, 1969.

[9] Sperlich, W. B. Namakir: A description of a Central Vanuatu language. PhD thesis, University of Auckland, Auckland, New Zealand, 1991.

[10] Lacrampe, S. Lelepa: Topics in the grammar of a Vanuatu language. PhD thesis, Australian National University, Canberra, Australia, 2014.

[11] Capell, A. The Nguna-Efate dialects, 1930-1980. http://catalog.paradisec.org.au/repository/AC2/VNEFAT11.

[12] Gordon, M. K. and Roettger, T. "Acoustic correlates of word stress: A cross-linguistic survey", Linguistics Vanguard, 3(1):1–11, 2017.

[13] Guy, J. B. M. A grammar of the northern dialect of Sakao. Number 33 in Series B. Pacific Linguistics, 1974.

[14] Jauncey, D. G. Tamambo, the language of west Malo, Vanuatu. Pacific Linguistics, 2011.

[15] Lynch, J. A grammar of Anejom̃. Pacific Linguistics, 2000.

[16] Lunden, A., Campbell, J., Hutchens, M., and Kalivoda, N. "Vowel-length contrasts and phonetic cues to stress: An investigation of their relation", Phonology, 34:565–580, 2017.

[17] Billington, R. Rosey Billington Nafsan materials (BR1), Digital collection managed by PARADISEC, 2017.

[18] Boersma, P. and Weenink, D. Praat, 2018. Version 6.0.40.

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

[20] Winkelmann, R., Harrington, J., and Jänsch, K. "EMU-SDMS: Advanced speech database management and analysis in R", Computer Speech & Language, 45:392–410, 2017.

[21] R Core Team. R: A language and environment for statistical computing, 2018. Version 3.4.2.

[22] Winkelmann, R., Jänsch, K., Cassidy, S., and Harrington, J. emuR: Main package of the EMU Speech Database Management System, 2018. R package, Version 1.0.0.

[23] 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.

[24] Beckman, M. Stress and non-stress accent. Foris Publications, 1986.

Investigating *word* prominence in Drehu

Catalina Torres, Janet Fletcher, Gillian Wigglesworth

The University of Melbourne, ARC Centre of Excellence for the Dynamics of Language

catalinat@unimelb.edu.au

Abstract

This study investigates the realization of informational focus in Drehu, an Oceanic language from New Caledonia. Stress in Drehu has informally been described as being demarcative and always falling on the first syllable of words. Our analysis of post-lexical accentuation shows a tendency for salient cues to be realized on a phrasal level. Results show a preference for marking the right edge with longer acoustic duration of final syllables and more extreme pitch movements. This evidence stands in contrast with the stress pattern reported in the literature and suggests a more detailed investigation of stress realization in Drehu is needed.

Index Terms: prominence marking, stress, Oceanic languages, Melanesia, Drehu, bilingual.

1. Introduction

Drehu is an Oceanic language from Lifou, New Caledonia, and the language with the largest number of speakers [1] in the archipel. According to the 2009 census [2], Lifou counts around 8600 inhabitants from which approximately 5500 are Drehu speakers older than 14 years old. Protestant missionaries developed the first writing system [3], which is in its majority still used for Drehu language class in primary, high school, and religious Sunday school. New Caledonia is a French overseas territory and the education system on the island follows the French metropolitan model. This means that apart from the optional Drehu language class all other subjects are taught in French. Today almost all speakers, especially younger generations, are bilinguals of French and Drehu.

Regarding the phonology of the language, it has been established that Drehu has a 30 consonants system including stops, nasals, fricatives, laterals, and approximants. Further, the inventory for vowels includes short and long vowels, and the language's phonotactics allow a syllabic structure with V, VC, CV, CVC and VV, VVC, CVV, CVVC [4]. Drehu has informally been described as a stress language with a word prosodic system [5, 6]. Stress is classified as demarcative, marking out word edges, and is not weight sensitive. According to [5], stress (accent d'intensité) always falls on the first syllable of a word, pëkö [ˈpɛkʌ] (*none, there is nothing*), fifikë [ˈfifikɛ] (*toy*). In word derivation, when words obtain a prefix, the stress pattern remains and stress shifts to the inserted first syllable e.g.: malan [ˈmalan] (*to fall*) vs. amalan [ˈamalan] (CAUS-*fall*). Compound words behave in the same way meaning that stress always shifts to the first syllable of the word. Finally, [6] proposes secondary stress in polysyllabic words, with it always falling on the third syllable [ˈamaˌlan (CAUS-*fall*)]. Within the Oceanic languages Drehu is classified as a language of the Southern Melanesian linkage and belongs to the Loyalty Islands family [7]. Drehu, Iaai, and Negone are closely related languages (Loyalty Islands) and were described as having similar phonological systems. They share several properties, like a rich vowel system that includes a length distinction, a lack of weight sensitivity regarding stress, which in turn is described as demarcative and fixed. Similarly, in Drehu, Iaai, and Nengone the rhythm type was analysed as trochaic [6, 8, 9, 10]. However, there are no phonetic studies further investigating the acoustics of rhythm or prosody in these three languages.

In this study the prosodic and phonetic realization of *word* prominence under informational focus is investigated in order to see whether the patterns recorded in the literature are born out in this pragmatic context. Our analysis is couched in the Autosegmental Metrical theory and aims at providing a first exploratory analysis of prosody in Drehu. According to intonational phonology [11], tonal events are defined through relative contrast from one tone to another. This contrast is perceived in relative tonal height and differs depending on the speaker's tone range. *High* tones are denoted with an (H) and *low* tones with an (L). Tonal events can be simple monotonal targets (H and L) but they can also be complex and bitonal (HL) falling or (LH) rising tones. In languages that bear stress, like Germanic languages, the stressed syllable acts as nuclear accent when in focus [12, 13]. Duration and fundamental frequency will be investigated to examine how they cue focal prominence and whether or not there is a specific set of acoustic parameters that are localised on particular syllables in the focal word. Regarding F0 it has also been pointed out that a nuclear accent in focus isn't necessarily realized with an increase in F0 but rather that a tonal event will accompany the syllable where it is realized. Hence, we are first of all interested in the tonal patterns as well as tonal movements found in focused constituents, and secondly in the acoustic measurements of F0 that accompany these tonal events.

2. Materials and Method

2.1. Participants

Four female speakers (age 29 - 47) were recorded in Lifou. Participants responded to a linguistic questionnaire similar to the Bilingual Language Profile [14]. All reported they acquired French and Drehu during childhood (starting at no later than 7 years with either language), were schooled in French, and had varying degrees of school instruction in Drehu (0 to 10 years). Participants were not only literate in French but also in Drehu. Additionally, they work in the local community in professions that require them to speak in the two languages (e.g. librarian, secretary).

2.2. Materials

Elicitation materials consist of three carrier phrases, in which 56 target words consisting of 2, 3, 4 and 5 syllables were inserted. Target words only contained short vowels, and had varying syl-

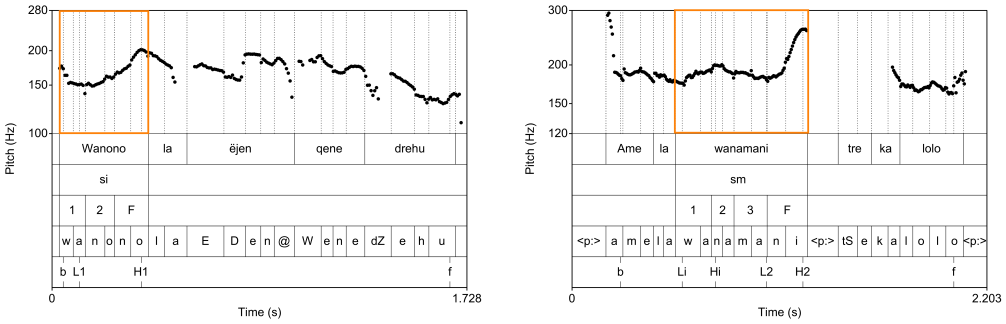


Figure 1: *The plot illustrates F0 contours and annotations of sentences containing the token 'wanono' (grey shrike) in sentence initial (si), and 'wanamani' (water drop) in sentence medial position (sm) with subsequent pause. The labels L1H1 indicate there is only one rise within the token. LiHi indicate there is one initial rise, and L2H2 indicate there is a second rise within the token.*

lable structure at the beginning or end of the word (V- CVC), but only CV syllables were included in the analysis for duration. Target tokens that included stop consonants or fricatives that did not allow reliable labelling of the F0 curve were not included in the analysis of intonation either. The position of the target phrase in the sentence was initial (*si*), medial (*sm*), and final (*sf*). Example (1) shows the carrier phrase used in sentence medial position. Prior to elicitation, the materials were checked for comprehensibility and suitability with a native speaker of Drehu.

(1) Ame la ___ tre ka lolo
ame la tʃe ka lolo
PRS1 ART ___ PRS2 STAT beautiful
'This ___ is beautiful/good'

2.3. Procedure

Recordings were made with a Zoom H6 in a quite room of a community centre. Materials were provided using powerpoint and partcipants had a training phase prior to recording. Sound files were manually transcribed, force aligned in WebMAUS, using a parameter model based on SAMPA [15], and then manually corrected in Praat [16]. Two acoustic parameters were analysed in this study, first duration, and second fundamental frequency. As exemplified in Figure 1, the target tokens, position, syllables and phones were marked. Subsequently, Tones were also manually annotated. A hierarchical database was constructed using the EMU Speech Database Management System [17]. It included tiers for the Tones, phonemic segments, syllables, target token position, and words. The acoustic and durational characteristics of vowel and word tokens produced in the target words were queried and analysed using the emuR package in R [18, 19].

2.4. Analysis

The experiment included three positions for the target tokens: sentence initial (*si*), sentence medial (*sm*), and sentence final (*sf*). Lifou speakers tended to insert pauses between the target tokens and the subsequent stream of speech. Pauses were coded and tokens followed by a pause were not included in statistical analyses. Due to the potential variability in duration of the onset consonants, and to allow for a more consistent comparison, the duration of all vowels in CV syllables was measured and analysed. Additionally, a broad ToBI-style annotation [20, 21] was used to investigate intonation in Drehu. The tonal targets were marked with L for low and H for high tones. Every subsequent tone identified in a token was also marked with an additional number. Figure 1 shows the points marked for tokens in *si*, and *sm* positions.

2.5. Duration hypothesis

As in [22] identified, duration has been recognized as the most frequent correlate of stress in a series of languages. Similarly, duration has been found to correlate with focus marking [23, 24]. Therefore, this parameter is the first one investigated in our study. Recall, stress in Drehu has been described as word initial, regardless of the number of syllables or morphological modifications, like affixation, adding additional syllables. As mentioned earlier, it has been claimed Drehu shows no weight sensitivity, meaning that also light syllables and short vowels can attract stress. Hence, we hypothesise that the first syllable and vowel contained in this syllable will show a greater duration in comparison to all other syllables in the word. Measurements of vowel duration for each CV syllable, in words of different lengths, were taken and then fitted into a linear mixed effects model in order to be compared. The model included 959 observations and included position of syllable and position in carrier sentence as fixed factors, plus speaker, word, and vowel as random factors. We used the step function to arrive at a final model and to obtain significance values, and used post-hoc Bonferroni correction to confirm significance of any interactions. All statistical analyses were carried out in R [19], using lme4 [25].

Table 1: *Identified tonal patterns, only in sentence initial and medial positions, Total = 205.*

Pattern	**LH**	**LHLH**	**LLH**	**HLH**	**LHL**	**HL**
Count	122	38	24	16	2	3
%	59.5	18.5	11.8	7.8	1	1.4

2.6. Intonation hypothesis

Similar to duration, F0 (fundamental frequency) has been recognized as the second most frequent acoustic correlate of stress showing most commonly (but not exclusively) greater F0 values on stressed syllables [22]. In many stress languages that have been previously analysed in an AM framework, the *nu-*

clear accent is a prominence lending pitch movement or target -i.e. a pitch accent, that is associated with the most prominent (stressed) syllable within the segmental string. If stress is word initial in Drehu, we hypothesise that under conditions of informational focus, this will also be the site of a pitch accent or major pitch event. Tonal targets were labelled and F0 values were extracted using Emu-R [18] in order to identify different tonal patterns of the experimental tokens. Additionally, measurements for F0 were taken for tonal targets (L and H Tones) found in tonal movements within tokens. Due to the nature of our data we examined rises and sought to find out if there was a difference in marking focus depending on where the rise was placed. We measured pitch range of rises as the difference in Hz between a H tone and its preceding L tone (See Figure 1). These values were fitted into a maximally specified linear mixed effects model which included 186 observations. The model included position of rise within token (LiHi, L1H1 or L2H2) as a fixed factor, and speaker as random factor. We used the step function to arrive at a final model and to obtain significance values, and used post-hoc Bonferroni correction to confirm significance of any interactions. This analysis was restricted to tokens in *si* and *sm* positions since the right edge of tokens followed by a pause or in *sf* position could be related to the marking of a higher prosodic level.

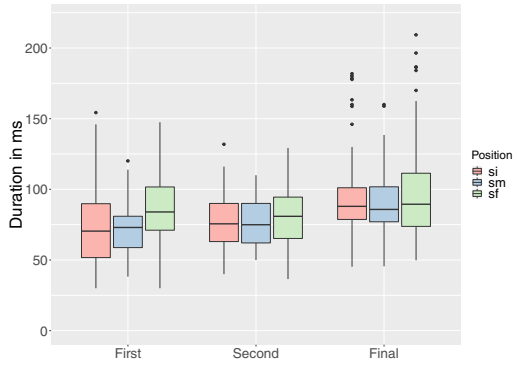


Figure 2: *The plot shows duration in milliseconds of vowels in CV structure of tokens containing two to three syllables, in three different positions: sentence initial (si), sentence medial (sm), and sentence final (sf) positions.*

3. Results

3.1. Duration

Figures 2 and 3 show the duration in milliseconds for vowels in CV syllables, and position of the syllable. Statistical analyses show that in the three different positions included in the experiment the vowel of the final syllable is significantly longer than the first vowel (t = - 8.39, p<0.001) and than all other vowels in preceding syllables. Hence, the first syllable of the word does not contain the longest vowel in the token. Additionally, there is no significant difference between vowels in initial versus medial syllables. These duration results do not provide evidence of word initial stress.

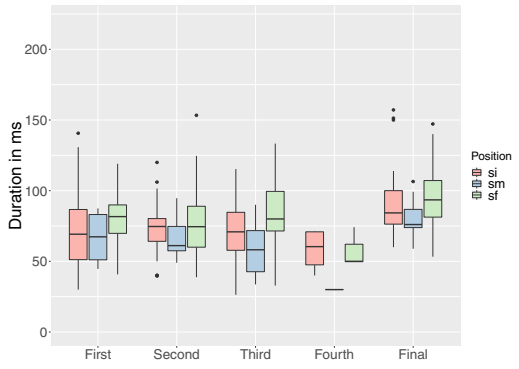


Figure 3: *The plot shows duration in milliseconds of vowels in CV structure of tokens containing four to five syllables, in three different positions: sentence initial (si), sentence medial (sm), and sentence final (sf) positions.*

3.2. Intonation

Table 1 shows the six different intonation patterns identified in our corpus. Four of these patterns represent 97.6% of all occurrences, and are constituted or end on a rising tone (LH). The most frequent pattern found (59,5%) was a rising tone with a H peak towards the right edge of the constituent. Rising tones were found to be either the tonal pattern of a whole constituent or a tonal movement within or at the right boundary of the constituent. Figure 4 shows a comparison of the three types of rises observed in our data: L1H1 tones that spread over the whole constituent; L2H2 tones occurring at the right edge; LiHi tones happening internally, prior to L2H2 (See also Figure 1). In tokens were the initial rise (LiHi) was found, 68% of the words had three or more syllables. There was a statistically significant difference in pitch range depending on the type of rise, with L1H1 (Est. 17 $\pm$ 5 Hz, p<0.002) and L2H2 (Est. 17 $\pm$ 6 Hz, p<0.002) showing greater pitch range expansion when compared to LiHi.

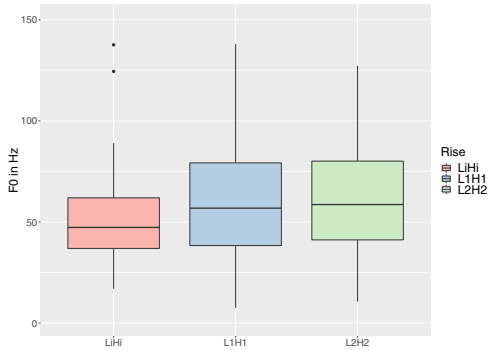


Figure 4: *The plot shows differences of pitch range in Hz of three different rising tones.*

4. Discussion and Conclusion

This study sought to provide a first evaluation of the phonetic marking of informational focus in Drehu. Our hypothesis re-

garding acoustic correlates of focus were confirmed by this data since duration and F0 proved to cue focal prominence consistently at the right edge of the constituent. Previous descriptions of Drehu phonology claimed that the language has initial word stress. In view of these analyses of Drehu as a stress language, we expected the first syllable of target tokens to constitute the location of a nuclear pitch accent in a word insertion task where the tokens were in informational focus. As duration and fundamental frequency have been widely claimed to be major correlates of stress and focus marking, we predicted the first syllable of the word should show longer duration values and a strong prominence-lending tonal movement. Contrary to our predictions, it was found that vowel duration was significantly longer on the *final* syllable of target tokens. Similarly, a rise was often observed, which in 97.6% of the cases ended on a peak at the right edge of the constituent. Measurements of F0 and a comparison of the pitch range of three different types of rises revealed that rises going across the whole constituent (L1H1) or rises placed directly prior to the right boundary (L2H2) showed significantly greater expansion than rises placed within the constituent (LiHi). Presumably, this word internal rise led to earlier interpretations of stress being placed on the first syllable of words. The results reported in our study show consistent tonal and durational marking of the right edge of constituents in focus. As the target tokens measured in our study are in informational focus, we can only interpret our findings in terms of the marking of focused constituents in Drehu. However, the lack of any durational or tonal cues associated with word-initial syllables calls for a further investigation, and a revaluation of previous stress analyses of Drehu, and perhaps also for other languages from the Loyalty Islands family. Finally, a more detailed study of prominence marking, for instance in non-focal words seems desirable in order to provide a more advanced description of Drehu prosodic phonology.

5. Acknowledgements

Special thank to the participants who made this study possible. This research was conducted with support from the ARC Centre of Excellence for the Dynamics of Language (Project ID: CE140100041).

6. References

[1] J. Vernaudon, "Linguistic Ideologies: Teaching Oceanic Languages in French Polynesia and New Caledonia," *The Contemporary Pacific*, vol. 27, no. 2, pp. 433–462, 2015.

[2] "Recensement général de la population," 2009.

[3] P. F. Magulué, "Les teachers du Pacifique au XIXe siècle ou l'émergence d'une nouvelle élite océanienne entre tradition et modernité," *Histoire et missions chrétiennes*, no. 4, pp. 139–156, 2011.

[4] C. Moyse-Faurie, *Le drehu, langue de Lifou (Iles Loyauté). Phonologie, morphologie, syntaxe*. Langues et Cultures du Pacifique Ivry, 1983.

[5] M.-H. Lenormand, "La phonologie du mot en lifou (Iles Loyalty)," *Journal de la Société des Océanistes*, vol. 10, no. 10, pp. 91–109, 1954.

[6] D. T. Tryon, *Dehu grammar*. Australian National University, 1968.

[7] T. Crowley, J. Lynch, and M. Ross, *The Oceanic languages*. Routledge, 2013.

[8] D. Tryon, *Iai Grammar*, ser. Pacific Linguistics, Series B. Canberra: Australian National University, 1968, vol. 8.

[9] D. T. Tryon, *Nengone Grammar*, ser. Pacific Linguistics, Series B. Canberra: Australian National University, 1967, vol. 6.

[10] D. T. Tryon and M.-J. Dubois, "Nengone dictionary. Part I, Nengone-English," in *Pacific Linguistics*. ERIC, 1969, vol. Series C-Books, No. 9.

[11] D. R. Ladd, *Intonational phonology*. Cambridge University Press, 2008.

[12] C. Gussenhoven, "Types of focus in English," in *Topic and focus: Cross-linguistic perspectives on meaning and intonation*, C. Lee and M. Gordon, Eds. Springer, 2007, pp. 83–100.

[13] S. Baumann, M. Grice, and S. Steindamm, "Prosodic marking of focus domains-categorical or gradient," in *Proceedings of speech prosody*, 2006, pp. 301–304.

[14] L. M. Gertken, M. Amengual, and D. Birdsong, "Assessing language dominance with the bilingual language profile," *Measuring L2 proficiency: Perspectives from SLA*, pp. 208–225, 2014.

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

[16] P. Boersma and D. Weenink, "Praat: doing phonetics by computer (version 6.0.26)[computer program]. Retrieved November 2, 2017," 2017.

[17] R. Winkelmann, J. Harrington, and K. Jänsch, "EMU-SDMS: Advanced speech database management and analysis in R," *Computer Speech & Language*, vol. 45, pp. 392–410, 2017.

[18] 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.

[19] 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/

[20] J. F. Pitrelli, M. E. Beckman, and J. Hirschberg, "Evaluation of prosodic transcription labeling reliability in the tobi framework," in *Third International Conference on Spoken Language Processing*, 1994.

[21] M. E. Beckman and J. Hirschberg, "The ToBI annotation conventions," *Ohio State University*, 1994.

[22] M. Gordon and T. Roettger, "Acoustic correlates of word stress: A cross-linguistic survey," *Linguistics Vanguard*, vol. 3, no. 1, 2017.

[23] F. Kügler, "The role of duration as a phonetic correlate of focus," in *Proceedings of the Speech Prosody 2008 Conference*. Editora RG/CNPq Campinas, Brazil, 2008, pp. 591–594.

[24] S.-A. Jun and C. Fougeron, "Realizations of accentual phrase in French intonation," *Probus*, vol. 14, no. 1, pp. 147–172, 2002.

[25] 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.

Recursive forced alignment: A test on a minority language

Simon Gonzalez[1], Catherine E. Travis[1], James Grama[1], Danielle Barth[1], Sunkulp Ananthanarayan[2]

[1] ARC Centre of Excellence for the Dynamics of Language, Australian National University
[2] University of Texas at Austin

simon.gonzalez@anu.edu.au, catherine.travis@anu.edu.au, james.grama@anu.edu.au, danielle.barth@anu.edu.au, sunny.a@utexas.edu

Abstract

We compare recursive and linear approaches to force-aligned data from Matukar Panau, an endangered language of Papua New Guinea. Data were force aligned with the train/align procedure in the Montreal Forced Aligner. Using manual alignments produced by a trained phonetician as a benchmark, the recursive approach was found to outperform the linear approach. The recursive approach produced alignments that overlapped more with those made by human coders, and resulted in fewer fluctuations in both Overlap Rate and Error Rate. We conclude that a recursive approach enhances the quality of automated alignment of languages lacking a pre-existing acoustic model.

Index Terms: forced alignment, accuracy, robustness, recursion, minority language

1. Introduction

Forced alignment is increasingly prevalent in phonetic research, as it dramatically increases the speed of achieving analysable data, and therefore the number of tokens that can be examined phonetically [1, 2]. Forced alignment has primarily been applied to major world languages with fully established acoustic models (in particular, English) [3, 4]. Thus, there exists a significant gap between the forced alignment resources available for minority languages, and those available for major languages. Minority languages have been force aligned using acoustic models from major languages, with varying success [5-8]. This approach is less than ideal because of the reliance it places on matching phonemic inventories and orthographic systems of often completely unrelated languages. We argue that current forced-alignment programs with a train/align procedure offer the means to effectively align data from languages which lack acoustic models, and, by applying a recursive approach, this is so even in the absence of large amounts of speech data with which to work.

A number of factors that affect the accuracy of forced alignment have been presented in the literature to date. Focus has been placed on best practices for addressing transcription errors [9], the impact of long pauses and noisy environments [10], as well as the optimal number of speakers and type of data for successful alignment [3]. Previous studies have also found that alignment accuracy tends to reach a ceiling, after which point additional data does not significantly improve the alignment. One study on spontaneous spoken English found that this ceiling was reached at five minutes of transcribed speech [3]; another, on read French data, exhibited a ceiling effect at two minutes [11]. Instead of increasing alignment quality, in some cases additional data was associated with in poorer alignment accuracy [3].

One of the more powerful tools in forced alignment is the train/align method. While some forced alignment works on the basis of a pre-existing acoustic model, with the train/align method, an acoustic model is created on the basis of the data input to the program, and that model is then applied to the forced alignment of the same input data [11]. This procedure has been successfully used to force align minority languages without established acoustic models [12, 13].

In working with minority, and under-resourced, languages, there may be limited data available, and thus maximal use must be made of the data that is available in order to build an acoustic model from scratch. The standard treatment of force-aligned data follows a linear approach, whereby the data is examined only once prior to creating a model. In contrast, in a recursive approach, the data is examined several times in different stages, and at every new stage, the algorithm learns from the previous stage and adapts accordingly [10]. It is therefore particularly valuable for working with small datasets.

This paper compares the application of a linear vs. recursive approach to force-aligned data from Matukar Panau, a minority language of Papua New Guinea, with no pre-existing acoustic model and with a moderately sized speech corpus [14, 15]. We demonstrate that the recursive approach yields very high quality alignment, and suggest that applying a recursive approach facilitates high quality forced alignments of under-described languages.

2. Methodology

2.1. Montreal Forced Aligner

The aligner chosen for this study was the Montreal Forced Aligner (MFA) [4]. MFA has been demonstrated to be more accurate than FAVE [16], MAUS [17] and Prosody lab-Aligner [18], and marginally more accurate than the train/align procedure in LaBB-CAT [19]. (See [20] for a comparison of these aligners.) One key difference is that these other forced aligners use the HTK toolkit, while MFA uses the Kaldi toolkit, which employs triphone acoustic models to better capture variability in phone realisations. Another is that MFA (like LaBB-CAT) allows for the application of the train/align method, facilitating extension to languages lacking an acoustic model.

2.2. Matukar Panau Speech Data

Matukar Panau is an endangered Oceanic language spoken by around 300 people in Madang Province, Papua New Guinea, in the village Matukar and hamlet Surumarang.

Documentation for this language is ongoing [14, 15]. It is an agglutinating, non-tonal language with 17 consonants, a small vowel inventory, and a fairly transparent orthography. There is no existing acoustic model of the language.

The data for this study come from a corpus of sixty short recordings of monologic narratives produced by 36 native speakers of Matukar Panau. The narratives were transcribed by a trained linguist in conjunction with six trained, semi-speakers of Matukar Panau (native speakers of Tok Pisin who have familiarity with Matukar Panau). Transcription was done at the utterance level, using a phonemically transparent orthography. For this study, we worked with 3.75 hours (or 225 minutes) of transcribed speech. We built a dictionary to map the 2,468 word types that occur in this sub-corpus to their phonemic representations, and force aligned the audio files using the train/align procedure in MFA. Data were then prepared following both a recursive and linear approach, as described below.

2.3. Data Recursion

How much data is required for effective forced alignment when working with a language with no acoustic model? To test the quality of forced alignment with different quantities of data, we needed to create subsets of the data of different lengths. To control for speaker effects, we had to include multiple speakers in each subset. Thus, a script was written to create a TextGrid file in Praat [21] which separated all files into increments of 30 seconds. The starting point was one minute from each of four files representing four speakers. From there, we created five-minute iterations by drawing 30-second increments from each file, until the file's duration was exhausted, at which point we drew a 30-second increment from a new file. We increased the data being aligned by five-minute iterations, until we reached the maximum duration of transcribed speech (225 minutes). The 30-second increments ensure that speakers were equally represented at each five-minute iteration, controlling for speaker effects at each step. The five-minute iterations allow for the quality of the forced alignment with different amounts of data to be compared, to identify the point at which alignment quality is optimised. This process resulted in a mean of 487 word tokens and 2,444 segments per five-minute iteration.

Two datasets were prepared from the forced alignment output: a *linear dataset* and a *recursive dataset*. Figure 1 provides a representation of the difference between the two. Linear processing is the default for forced alignment. To prepare the linear dataset for this study, the force-aligned boundaries were *reset* at each five-minute iteration. In contrast, the force-aligned boundaries for the recursive dataset were *adjusted*. That is, for each iteration, the alignment was recalculated based on the information from the current iteration and from previous iterations, utilising an algorithm that was written for this process. This methodology is adapted from [10], which utilised a recursive algorithm to improve forced alignment in long audio segments. A recursive algorithm works by inspecting the data multiple times; at every new iteration, new information is added, re-evaluated, fed back into the existing information from previous iterations, and then applied at the next iteration. This output then serves as the basis for analysis.

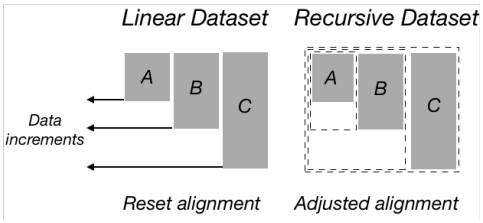


Figure 1 *Alignment process according to linear and recursive datasets*

3. Measures of alignment quality

The standard for determining the quality of forced alignment is comparison with a human benchmark. One method available for doing this is a comparison between boundaries placed by forced alignment vs. boundaries placed by a trained phonetician [3, 22]. Following the same protocol as that applied in previous work, we selected two speakers (one male and one female), and manually corrected the automatic alignment of the first 60 seconds of the file of each speaker. These 549 segments (261 consonants and 288 vowels) serve as the benchmark against which the automatic alignments are compared.

Two quality measurements are employed in this study: *accuracy* and *robustness* [cf., 23]. *Accuracy* was operationalized as the time difference between the placement of a boundary as the result of forced alignment vs. the human benchmark. *Robustness* is the rate of alignment error based on a specified boundary threshold, here set at 20 *ms*, following [3]. Any force-aligned boundary placed greater than 20 *ms* from the human benchmark is classified as an alignment error.

These measurements provide different indications of the quality of the forced alignment. An alignment may have high accuracy but not be robust if there are a large number of alignments that occur just beyond the 20 *ms* threshold, but close to that threshold. On the other hand, an alignment can have robust alignments with low accuracy if there are few alignments beyond 20 *ms*, but those that are beyond 20 *ms* are at a high degree of distance from the benchmark. Together, these measures provide a strong indication of the overall quality of the resulting alignments.

3.1. Accuracy Measurement

For accuracy, we calculated Overlap Rate (OvR) [3, 23], that is, the proportion of overlap between the intervals established by the human coder and the intervals established by the forced aligner. Greater overlap is associated with greater accuracy. The time representation is shown in Figure 2.

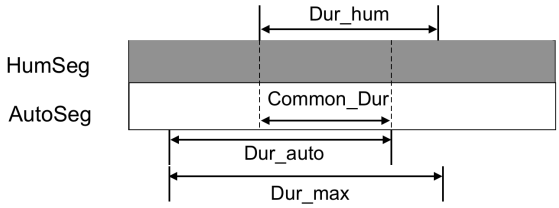


Figure 2: *Representation of Overlap Rate*

Common_Dur is the time shared by the interval created by the human coder (*Dur_hum*), and that created automatically, by the forced aligner (*Dur_auto*). This is measured as a

proportion of the duration from the earliest onset and latest offset boundary of the two intervals (*Dur_max*). The Overlap Rate was calculated for both linear and recursive datasets.

3.2. Robustness Measurement

For robustness, we are interested in the proportion of boundaries that lie beyond a pre-determined threshold, here 20 ms. Following [13], this was calculated on the basis of the difference between the midpoint of the manually created interval and the force-aligned interval. Figure 3 shows a hypothetical midpoint of an interval produced by a human coder, and two hypothetical midpoints from distinct forced alignments. If a force-aligned midpoint falls within 20 ms of a manual midpoint (as for (a)), it is considered a non-error; if it is at a greater distance (as for (b)), it is considered an error. The Error Rate is the ratio of total number of errors to non-errors. As a further measure of robustness, we calculated the mean distance from the manual midpoint of the error tokens. Higher mean distances correspond to less robust tokens, thus less reliable alignments. The two measures of robustness were calculated for both linear and recursive datasets.

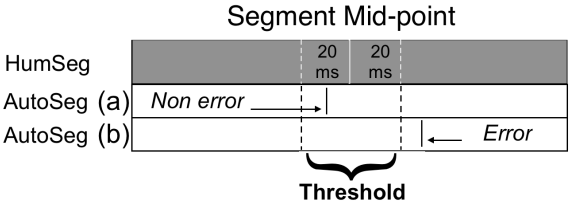


Figure 3: *Representation of errors, using a 20 ms threshold*

4. Results

4.1. Accuracy

Figure 4 compares the Overlap Rate for the linear and recursive datasets across iterations for the two speakers who were manually aligned. The greater accuracy for the recursive dataset, captured with the solid line, can be seen in three ways (all of which hold for each speaker).

First, the overall mean Overlap Rate is higher in the recursive than the linear dataset. Second, the recursive dataset exhibits fewer fluctuations than the linear dataset, suggesting that a recursive approach smooths out major alignment errors. And third, in the recursive dataset, there is a marked increase in Overlap Rate up to 35 minutes of data; beyond this point, the increase is more gradual. In comparison, the linear dataset retains significant fluctuations throughout, though they become less marked from approximately 125 minutes. Thus, the recursive dataset improves more rapidly, follows a steadier trajectory, and overlaps more with alignments placed by a human coder than the linear dataset.

Evidence of the quality of the alignment can be seen by comparing these results with the findings of [3] for English. The Overlap Rate of 0.67 attained at 35 mins in the recursive dataset here is at the upper end of the range reported in [3]; but in [3], this was attained earlier, with just five minutes of data.

4.2. Robustness

The overall Error Rate, that is, the proportion of midpoints determined automatically that occurred at a distance of greater than 20 *ms* from the human benchmark, was very similar across the two approaches (recursive dataset = 21.5%, linear dataset = 23.2%). Overall mean distance from the human benchmark for tokens classified as errors was also similar (recursive dataset = 93.6 *ms*, linear dataset = 94.9 *ms*). Thus, according to this measure of robustness, the recursive dataset produces only marginally better results.

However, mean distances across data iterations differ. Figure 5 shows the mean distance from the benchmark across data iterations for tokens classified as errors. Here we see that, while the recursive dataset exhibits a steadier trajectory throughout, the linear approach is characterized by heavy fluctuation, with pronounced differences between peaks and troughs. As with Overlap Rate, the recursive approach seems to be able to soften the impact of errors more efficiently than the linear approach.

We also see here that both linear and recursive datasets show stabilisation at approximately 35 minutes (a similar point at which improvement in accuracy according to Overlap Rate begins to diminish). And both datasets show a decrease in robustness at the latter stages of data iteration—the linear dataset exhibits striking variability, and the recursive dataset shows a gradual increase in mean distance. One possible explanation is that the greater number of speakers included at these latter stages may result in more variability, and we leave this for future exploration.

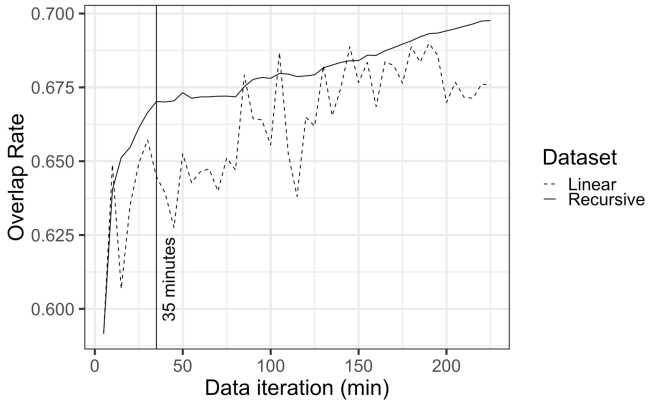


Figure 4: *Overlap Rate across data iterations: Linear and Recursive datasets*

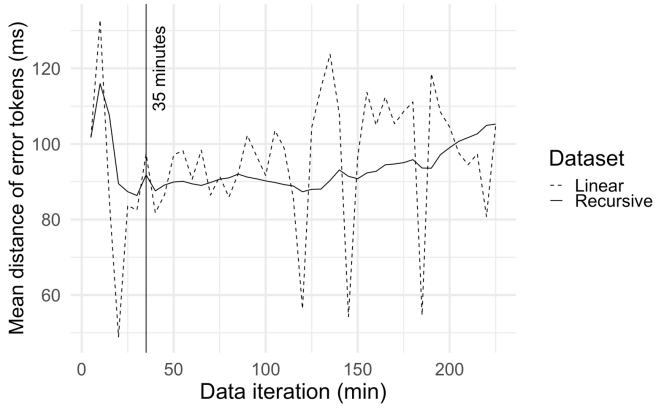


Figure 5: *Mean distance of error tokens across data iterations: Linear and Recursive datasets*

5. Discussion

In this study, we applied a recursive approach to forced aligned data from Matukar Panau, a minority language of Papua New Guinea lacking an acoustic model. To do this, we utilised the train/align procedure in MFA, and aligned the same data at different iterations, from 5 to 225 minutes. We then applied two approaches to prepare data for analysis: a linear approach, where alignment values are reset at each iteration, and a recursive approach, where alignment values are adjusted based on previous iterations. Results indicate that a recursive approach outperforms a traditional linear approach—forced alignments derived via recursion were more accurate, with a higher rate of overlap between manually and automatically placed boundaries. The recursive approach was also more robust than the linear approach in the sense that it was less susceptible to major alignment errors. This suggests that the algorithm learns as more data is processed, in line with observations that adjustments early on in alignment improve alignment in later stages [10]. In this way, the recursive approach may be able to protect alignment in sections of audio files (or whole audio files) that prove challenging to aligners. As the recursive approach employs an algorithm that is self-correcting, these mistakes can be adjusted for as the data is processed; the algorithm learns from these examples and this ultimately improves the alignment later in the data stream.

6. Conclusions

The recursive approach to the force-aligned data outperforms traditional linear implementations, and yields highly accurate alignment, even from a relatively small dataset—here, 35 minutes of transcribed speech was sufficient to achieve high quality alignment. This method expands the potential for large-scale phonetic and sociophonetic studies for under-resourced minority languages, and is a very promising step towards making available to minority languages tools that to date have been primarily utilised for work on major world languages. Future studies on languages of different types will further advance methods and the ability to obtain the best results for automated phonetic alignment.

7. Acknowledgements

We gratefully acknowledge support from an ARC Centre of Excellence for the Dynamics of Language Transdisciplinary & Innovation Grant (TIG952018), as well as the local transcription team (Justin Willie, Rudolf Raward, Amos Sangmei, Alfred Sangmei, Michael Balias and Zebedee Kreno), and expert consultant (Kadagoi Rawad Forepiso).

8. References

[1] Labov, W., Rosenfelder, I., and Fruehwald, J., “One hundred years of sound change in Philadelphia: Linear incrementation, reversal, and reanalysis,” Language, 89(1):30-65, 2013.

[2] Schiel, F. et al., "The Production of Speech Corpora," 2012.

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

[4] McAuliffe, M. et al., “Montreal Forced Aligner: Trainable text-speech alignment using Kaldi,” Proceedings of the 18th Conference of the International Speech Communication Association, 2017.

[5] Kempton, T., "Cross-language forced alignment to assist community-based linguistics for low resource languages," Paper presented at the 2nd Workshop on Computational Methods for Endangered Languages, ComputEL-2, Honolulu. 2017.

[6] Kempton, T., Moore, R. K., and Hain, T., “Cross-language phone recognition when the target language phoneme inventory is not known,” INTERSPEECH-2011:3165-3168, 2011.

[7] Kurtic, E. et al., “A corpus of spontaneous multi-party conversation in Bosnian Serbo-Croatian and British English,” Proceedings of the Eighth International Conference on Language Resources and Evaluation, LREC’12, Istanbul, Turkey, 1323-1327, 2012.

[8] Strunk, J., Schiel, F., and Seifart, F., “Untrained forced alignment of transcriptions and audio for language documentation corpora using WebMAUS,” Proceedings of the Ninth International Conference on Language Resources and Evaluation, LREC’14, Reykjavik, Iceland, 3940–3947, 2014.

[9] Bordel, G., Penagarikano, M., Rodriguez-Fuentes, L. J., and Varona, A., “A simple and efficient method to align very long speech signals to acoustically imperfect transcriptions,” INTERSPEECH-2012:1840-1843, 2012.

[10] Moreno, P. J., Joerg, C. F., Van Thong, J.-M., and Glickman, O., "A recursive algorithm for the forced alignment of very long audio segments," Paper presented at the 5th International Conference on Spoken Language Processing, Incorporating The 7th Australian International Speech Science and Technology Conference, Sydney Convention Centre, Australia. 1998.

[11] Brognaux, S., Roekhaut, S., Drugman, T., and Beaufort, R., “Automatic phone alignment: A comparison between speaker-independent models and models trained on the corpus to align,” Proceedings of the 8th International Conference on NLP, JapTAL:300–311, 2012.

[12] DiCanio, C. et al., “Assessing agreement level between forced alignment models with data from endangered language documentation corpora,” INTERSPEECH-2012:130-133, 2012.

[13] Coto-Solano, R., and Flores, S., “Comparison of two forced alignment aystems for aligning Bribri speech,” CLEI Electronic Journal, 20(1):2:1-2:13, 2017.

[14] Anderson, G. D. S., Barth, D., and Rawad Forepiso, K., “The Matukar Panau online talking dictionary: Collective elicitation and collaborative documentation,” Language documentation and cultural practices in the Austronesian world, Papers from 12-ICAL, Canberra, Australia, 111-126, 2015.

[15] Barth, D., “Matukar Panau Language Documentation (DGB1), Digital collection managed by PARADISEC. [Open Access] DOI: 10.4225/72/56E97A2420C64,” 2010.

[16] Rosenfelder, I. et al., "FAVE (Forced Alignment and Vowel Extraction) Program Suite v1.2.2 10.5281/zenodo.22281," 2014.

[17] Schiel, F., “Automatic phonetic transcription of non-prompted speech,” ICPhS-14:607-610, 1999.

[18] Gorman, K., Howell, J., and Wagner, M., “Prosody lab-aligner: A tool for forced alignment of laboratory speech,” Canadian Acoustics, 39(3):192-193, 2011.

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

[20] Gonzaléz, S., Grama, J., and Travis, C. E., “Comparing the accuracy of forced-aligners for sociolinguistic research,” Poster presented at CoEDL Fest, University of Melbourne (available at: https://cloudstor.aarnet.edu.au/plus/s/gyC6vuX5uvc5soG - pdfviewer), 2018.

[21] Boersma, F., J., and Weenink, D., Praat: Doing phonetics by computer [Computer Software] Amsterdam: Department of Language and Literature, University of Amsterdam. Retrieved from http://www.praat.org/, 2011.

[22] Cosi, P., Falavigna, D., and Omologo, M., “A preliminary statistical evaluation of manual and automatic segmentation discrepancies,” 1991.

[23] Paulo, S., and Oliveira, L. C., "Automatic phonetic alignment and its confidence measures," Advances in Natural Language Processing: 4th International Conference, EsTAL 2004, Luis Vicedo, J. et al., eds., 36-44, Berlin / Heidelberg: Springer-Verlag 2004.

Use of Uncertainty Propagation in Twin Model GPLDA for Short Duration Speaker Verification

Jianbo Ma[1,2], Vidhyasaharan Sethu[1], Eliathamby Ambikairajah[1,2], Kong Aik Lee[3]
[1]School of Electrical Engineering and Telecommunications, UNSW Sydney
[2]DATA61, CSIRO, Sydney, Australia
[3]Data Science Research Laboratories, NEC Corporation, Japan
jianbo.ma@unsw.edu.au

Abstract

In automatic speaker verification, uncertainty propagation in Gaussian Probabilistic Linear Discriminant Analysis (GPLDA) was proposed to take into account duration variability in utterances. Independently, the Twin Model GPLDA for speaker verification was developed where long utterances were used for speaker enrolment and the system was tested on short utterances. In this paper, we introduce the use of uncertainty propagation in the Twin Model GPLDA to further improve its modelling ability under duration mismatched conditions. This approach takes into account both, the differences in distributions of i-vectors from long and short duration utterances, and the uncertainty associated with each i-vector. The proposed approach was tested on the NIST SRE 2010 8CONV-10SEC condition as well as additional 5 and 3 seconds conditions. Experimental results show that the proposed approach leads to improved speaker verification performance, especially when there are a limited number of speakers in the training data.

Index Terms: uncertainty propagation, GPLDA, Twin Model GPLDA, short duration, automatic speaker verification

1. Introduction

Automatic speaker verification systems can be broadly categorised into one of two types: text-dependent and text-independent systems. In text-dependent systems, the lexical content of utterances are fixed (e.g., speaker-specific pass-phrase), which in turn allows text-dependent models to identify speakers given the phonetic context [1]. In contrast, speakers are free to speak any phrases and the system cannot rely on prior knowledge of fixed pass-phrases in the text-independent case [2]. While the development of text-independent systems is recognised as a more challenging task, text-independent systems also have a greater number of applications compared to text-dependent systems. Most state-of-the-art text-independent speaker verification systems are based on i-vector/GPLDA systems [3, 4]. Typically these systems utilise long utterances (e.g. 2 to 3 minute utterances) for both enrolment and testing. However, short duration speaker verification (i.e., short test utterances) would be significantly more practical. For example, access control usually requires machines to verify identities within a few seconds and a speaker verification system that requires a long utterance (several minutes) for verification would not be suitable. Since enrolment is carried out only once in an offline manner, it is still reasonable to assume that long utterances can be used for enrolment. Consequently this paper focusses on the use of long utterances for enrolment and short utterances for testing.

It has been shown that such duration mismatch between enrolment and test data can lead to significant degradation in performance [5], and this mismatch needs to be addressed in order to make short duration speaker verification a viable option. Some methods to tackle this problem have been previously proposed [5-8] under the assumption that i-vector is representative of a given utterance. For example, in [6], score domain compensation for duration mismatch was introduced. Most recently, the Twin Model GPLDA (TM-GPLDA) was proposed to model variations in distributions caused by duration mismatch [7]. However, given that the i-vector is a point estimate based on the posterior distribution of the latent variables in a total variability model, the uncertainty of this point estimation will become larger as the duration of utterance reduces [9]. Thus, the i-vector becomes less reliable when the utterance is short. A variation of the GPLDA model that takes into account this uncertainty of the i-vector was then proposed and shown to be effective [10, 11]. As illustrated in [5, 7], utterances with different duration (even from the same speaker) leads to differences in the distribution of i-vectors, which is addressed by the TM-GPLDA. On the other hand, the GPLDA with uncertainty propagation [10], which targets arbitrary durations, focuses on uncertainty in the distribution of latent variables.

From the observations above, it can be seen that a single model that explicitly takes into account both the uncertainty of i-vectors and the distribution mismatch between long and short durations would be beneficial and is the focus of this paper. The effectiveness of the proposed TM-GPLDA with uncertainty propagation was tested on the NIST SRE 2010 8CONV-10SEC condition and additional 5 and 3 seconds conditions.

2. Background

Given a set of i-vectors $\chi = \{\omega_{ij}; i = 1,2,\cdots,S; j = 1,2\cdots,J_i\}$, where ω_{ij} denotes the i-vector corresponding to the j^{th} utterance from the i^{th} speaker, GPLDA [12] decomposes them as:

$$\omega_{ij} = \mu + \Phi y_i + \varepsilon_{ij} \quad (1)$$

where Φ is a factor loading matrix, y_i is a vector of latent variables that follow the distribution $\mathcal{N}(0,\mathrm{I})$, and residual term ε_{ij} is assumed to follow Gaussian distribution with zero mean and a full covariance matrix denoted by Σ.

In contrast, the generative model of TM-GPLDA [7] can be written as:

$$\omega = \begin{cases} \mu_l + \Phi_l y + \varepsilon_l, & \text{for long utterances} \\ \mu_s + \Phi_s y + \varepsilon_s, & \text{for short utterances} \end{cases} \quad (2)$$

where ω denotes the i-vector; μ_l and μ_s are mean vectors for i-vector correspond to long and short utterances, respectively; Φ_l and Φ_s are the corresponding factor loading matrices; y is the vector of normally distributed latent variables and is

shared by all the utterances from the same speaker (for all long or short utterances), and ε_l and ε_s are residuals that vary across utterances and are assumed to be normally distributed with zero mean and covariance given by the matrices Σ_l and Σ_s for long and short utterances respectively. Note that the speaker label subscripts are omitted in this generative equation.

Finally, the generative model behind GPLDA with uncertainty propagation [10] is

$$\omega_{ij} = \mu + \Phi y_i + U_j x_j + \varepsilon_{ij} \tag{3}$$

where $U_j U_j^*$ is the Choleskey decomposition of posterior covariance matrix associated with corresponding i-vector and x_j is a hidden variable having a standard normal distribution. Other parameters are the same as those used in equation (1). Training and scoring formulae corresponding to these models can be found in [7, 10].

3. Uncertainty propagation in TM-GPLDA

In order to introduce uncertainty propagation, the generative model of TM-GPLDA is modified as:

$$\omega_{ij} = \begin{cases} \mu_l + \Phi_l y_i + U_{ij,l} x_{ij,l} + \varepsilon_{ij,l}, \text{for long utterances} \\ \mu_s + \Phi_s y_i + U_{ij,s} x_{ij,s} + \varepsilon_{ij,s}, \text{for short utterances} \end{cases} \tag{4}$$

where, l denotes long utterances class and s for short; μ_l, μ_s, Φ_l, Φ_s and y_i are as defined for the standard TM-GPLDA given in (2); the residual terms $\varepsilon_{ij,l}$ and $\varepsilon_{ij,s}$ follow $\mathcal{N}(0, \Sigma_l)$ and $\mathcal{N}(0, \Sigma_s)$ respectively; $x_{ij,l}$ and $x_{ij,s}$ are latent variables having standard normal distributions; $U_{ij,l}U_{ij,l}^*$ and $U_{ij,s}U_{ij,s}^*$are the Choleskey decompositions of posterior covariance matrices associated with corresponding i-vectors. The EM algorithm is used to estimate the model hyper-parameters and the formulations are developed in following sections.

3.1. Expectation step

In the E-step, posterior probabilities of latent variables are estimated. According to Bayes rule, the posterior distribution is proportional with the production of likelihood and prior, which is expressed as

$$p(y_i|\omega_i, \theta) \propto p(\omega_i|y_i, \theta) p(y_i) \tag{5}$$

where ω_i denotes the set of i-vectors from the i^{th}speaker. The likelihood term is expanded as:

$$\begin{aligned} p(\omega_i|y_i, \theta) &= p\left(\omega_i|y_i, x_{i1,l}, \dots, x_{iJ_i,l}, x_{i1,s}, \dots, x_{iJ_i,s}, \theta\right) \\ &= \prod_i \prod_j \prod_{k \in l,s} \int p\left(\omega_{ij,k}|y_i, x_{ij,k}, \theta\right) p\left(x_{ij,k}\right) dx_{ij,k} \end{aligned} \tag{6}$$

where

$$\begin{aligned} p\left(\omega_{ij,k}|y_i, x_{ij,k}, \theta\right) = \mathcal{N}\left(\omega_{ij,k}|\mu_k + \Phi_k y_i \right. \\ \left. + U_{ij,k} x_{ij,k}, \Sigma_k\right) \end{aligned} \tag{7}$$

and the prior distribution of latent variable x is a standard normal distribution. According to the convolution of Gaussian kernel [13], the integral in (6) is valued as $\mathcal{N}(\omega_{ij,k}|\mu_k + \Phi_k y_i, \Sigma_k + U_{ij,k}U_{ij,k}^*)$. Notice that the term $U_{ij,k}U_{ij,k}^*$ is now part of the residual covariance.

As in [14], we write the likelihood term as a stacked equation as follows:

$$\begin{bmatrix} \omega_{i1,l} \\ \dots \\ \omega_{iJ_{il},l} \\ \omega_{i1,s} \\ \dots \\ \omega_{iJ_{is},s} \end{bmatrix} = \begin{bmatrix} \mu_l \\ \dots \\ \mu_l \\ \mu_s \\ \dots \\ \mu_s \end{bmatrix} + \begin{bmatrix} \Phi_l \\ \dots \\ \Phi_l \\ \Phi_s \\ \dots \\ \Phi_s \end{bmatrix} y_i + \begin{bmatrix} \tilde{\varepsilon}_{i1,l} \\ \dots \\ \tilde{\varepsilon}_{iJ_{il},l} \\ \tilde{\varepsilon}_{i1,s} \\ \dots \\ \tilde{\varepsilon}_{iJ_{is},s} \end{bmatrix} \tag{3}$$

It then can be written as a Gaussian kernel $\mathcal{N}(\omega_i|\mu_i + \widetilde{\Phi}_i y_i, \tilde{\Sigma}_i)$, where

$$\omega_i = \begin{bmatrix} \omega_{i1,l} \\ \dots \\ \omega_{iJ_{il},l} \\ \omega_{i1,s} \\ \dots \\ \omega_{iJ_{is},s} \end{bmatrix}, \mu_i = \begin{bmatrix} \mu_l \\ \dots \\ \mu_l \\ \mu_s \\ \dots \\ \mu_s \end{bmatrix}, \widetilde{\Phi}_i = \begin{bmatrix} \Phi_l \\ \dots \\ \Phi_l \\ \Phi_s \\ \dots \\ \Phi_s \end{bmatrix}$$

$$\tilde{\Sigma}_i = \begin{bmatrix} \Sigma_l + U_{ij,l}U_{ij,l}^* & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \Sigma_s + U_{ij,s}U_{iJ_{is},s}^* \end{bmatrix}$$

The posterior probability is then proportional to a product of two Gaussian kernels, expressed as:

$$p(y_i|\omega_i, \theta) \propto \mathcal{N}\left(\omega_i|\mu_i + \widetilde{\Phi}_i y_i, \tilde{\Sigma}_i\right) \mathcal{N}(y_i|0, I) \tag{9}$$

Based on the fact that the outcome of product of two Gaussians is still a Gaussian, after some algebraic manipulations, the first and second moment of the posterior probability are expressed as

$$\begin{aligned} E(y_i) &= \left(I + \widetilde{\Phi}_i^* \tilde{\Sigma}_i^{-1} \widetilde{\Phi}_i\right)^{-1} \widetilde{\Phi}_i^* \tilde{\Sigma}_i^{-1} (\omega_i - \mu_i) \\ E(y_i y_i^*) &= \left(I + \widetilde{\Phi}_i^* \tilde{\Sigma}_i^{-1} \widetilde{\Phi}_i\right)^{-1} + E(y_i) E(y_i^*) \end{aligned} \tag{10}$$

Similarly, the first and second moment of the posterior probability of latent variable $x_{ij,k}$ are calculated as:

$$\begin{aligned} E\left(x_{ij,k}\right) &= \left[I + U_{ij,k}^* (\Phi_k \Phi_k^* + \right. \\ &\left. \Sigma_k)^{-1} U_{ij,k}\right]^{-1} U_{ij,k}^* (\Phi_k \Phi_k^* + \Sigma_k)^{-1} \left(\omega_{ij,k} - \mu_k\right) \\ E\left(x_{ij,k} x_{ij,k}^*\right) &= \left[I + U_{ij,k}^* (\Phi_k \Phi_k^* + \Sigma_k)^{-1} U_{ij,k}\right]^{-1} \\ &+ E\left(x_{ij,k}\right) E\left(x_{ij,k}^*\right) \end{aligned} \tag{11}$$

where $k \in \{l, s\}$.

3.2. Maximization step

The auxiliary function of EM algorithm is of the form:

$$Q(\theta, \theta_{old}) = \int p(H|\omega, \theta_{old}) log[p(\omega|H) p(H)] dH \tag{12}$$

where, H denotes the collected latent variables including y and x, and they are independent, ω denotes i-vectors, and θ_{old} denotes the model hyper-parameters from the previous iteration of the EM algorithm. It is regarded as a lower bound of log-likelihood of observable data and each step will increase the log-likelihood, leading to a local optimum of parameters. By observing the generative model, it can be found that given the latent variable, the observable variables are independent and parameters are associated with corresponding long or short class. This indicates that the auxiliary function is a linear combination of different classes. In the M-step, we optimize the auxiliary function:

$$\tilde{Q}(\theta, \theta_{old}) = \sum\nolimits_k Q(\theta_k, \theta_{old}) \tag{13}$$

where

$$Q(\theta_k, \theta_{old}) = \int p(H|\omega, \theta_{old}) log[p(\omega_k|H) p(H)] \, dH \tag{14}$$

where, ω_k denotes i-vectors from k class (long or short utterance). The auxiliary function is written as a summation of long and short class in terms of parameters, which means that these two sets of parameters can be updated separately. By setting the derivative with respect to each parameter to zero, and after some algebraic manipulations, the update equations can be written as:

$$\begin{aligned}\Phi_k &= \{\Sigma_{ij}[\omega_{ij,k} - \mu_k - U_{ij,k}E(x_{ij,k})]E(y_i)\}[\Sigma_{ij}E(y_i y_i^*)]^{-1} \\ \Sigma_k &= \frac{1}{\Sigma_i J_{i,k}}\{\Sigma_{ij}[(\omega_{ij,k} - \mu_k)(\omega_{ijk} - \mu_k)^* - [\Phi_k E(y_i) + \\ &\qquad U_{ij,k}E(x_{ij,k})](\omega_{ij,k} - \mu_k)]\}\end{aligned} \quad (15)$$

where $k \in \{l, s\}$.

3.3. Scoring

Based on the TM-GPLDA with uncertainty propagation model, given an enrolment i-vector ω_e and a test i-vector ω_t from a trial, the log-likelihood ratio between the hypothesis that the two i-vectors are from the same speaker (H_0) versus the hypothesis that they are from different speakers (H_1) is calculated as:

$$Score(\omega_e, \omega_t) = log\frac{p(\omega_e, \omega_t|H_0)}{p(\omega_e, \omega_t|H_1)} \quad (16)$$

where

$$p(\omega_e, \omega_t|H_0) = \int p(\omega_e, \omega_t|y, x_e, x_t)p(y)p(x_e)p(x_t)dydx_e dx_t$$

$$p(\omega_e, \omega_t|H_1) = \int p(\omega_e|y_e, x_e)p(y_e)p(x_e)dy_e dx_e \int p(\omega_t|y_t, x_t)p(y_t)p(x_t)dy_t dx_t$$

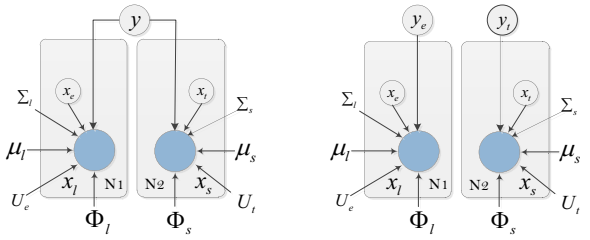


Figure 1: *(a) Hypothesis that test and enrolment i-vectors are from same speaker (share latent variables - y); (b) Hypothesis that test and enrolment i-vectors are from different speakers (distinct latent variables - y_e and y_t)*

Figure 1 shows the graphical models corresponding to the two hypotheses. The evaluation procedure is similar to the E-step. Likelihood terms are calculated as:

$$\begin{aligned}&p(\omega_e, \omega_t|H_0) \\ &= \mathcal{N}\left(\begin{bmatrix}\omega_e \\ \omega_t\end{bmatrix}|\begin{bmatrix}\mu_l \\ \mu_s\end{bmatrix}, \begin{bmatrix}\Sigma_l + \Phi_l\Phi_l^* + U_e U_e^* & \Phi_l\Phi_s^* \\ \Phi_s\Phi_l^* & \Sigma_s + \Phi_s\Phi_s^* + U_t U_t^*\end{bmatrix}\right) \\ &p(\omega_e, \omega_t|H_1) \\ &= \mathcal{N}\left(\begin{bmatrix}\omega_e \\ \omega_t\end{bmatrix}|\begin{bmatrix}\mu_l \\ \mu_s\end{bmatrix}, \begin{bmatrix}\Sigma_l + \Phi_l\Phi_l^* + U_e U_e^* & 0 \\ 0 & \Sigma_s + \Phi_s\Phi_s^* + U_t U_t^*\end{bmatrix}\right)\end{aligned} \quad (17)$$

3.4. Scaling covariance

From previous sections, it can be seen that the uncertainty of an i-vector is propagated into the covariance of the marginal distribution in the calculation of posterior probability and final score. This covariance consists of three terms, which are Σ_k, $\Phi_k\Phi_k^*$ and uncertainty of corresponding i-vector. As the uncertainty tends to be larger with decrease in duration, the covariance of latent variable in total variability model of short utterances may become overwhelming in hyper-parameters training phase, leading to poor modeling ability of GPLDA. To solve this problem, scaling factors are introduced into TM-GPLDA with uncertainty propagation model. The generative model is then revised as:

$$\omega_{ij} = \begin{cases}\mu_l + \Phi_l y_i + \lambda_l U_{ij,l}x_{ij,l} + \varepsilon_{ij,l}, & \text{for long utterances} \\ \mu_s + \Phi_s y_i + \lambda_s U_{ij,s}x_{ij,s} + \varepsilon_{ij,s}, & \text{for short utterances}\end{cases} \quad (18)$$

where λ_l and λ_s are scaling factors. The values are empirically determined by experiments. Our basic assumption is that the structure of covariance matrix for a given class (long or short) is important and scaling factors are introduced to balance the contributions from different uncertainties (uncertainty from single i-vector or from the residual in equation 4). The validity of this assumption is supported by [15], where an identical covariance can be shared by utterances that are of relatively same length. Instead of using same covariance, we relax it to have the same scaling factor. The scaling factors can be directly absorbed into the $U_{ij,k}$ term and the same modelling, training, scoring equations can be used.

4. Experiments and discussion

The 8CONV-10SEC task (condition 5) of the NIST SRE'10 [16] was chosen for the experiments. Two additional conditions were created by truncating the 10 seconds test utterances to 5 and 3 seconds (using the first 5 seconds and 3 seconds of each utterance).

The baseline system is an i-vector/GPLDA system. Standard MFCC features of 13 dimensions with their first and second derivatives were used in conjunction with a vector quantization model based voice activity detector [17] prior to feature warping [18]. Only female speakers were considered in the experiments reported in this paper and a gender-dependent universal background model (UBM) of 1024 Gaussian mixtures was created using utterances from female speakers from NIST SRE'04, 05, 06, 08, Switchboard II Part 1, 2, 3 and Switchboard Cellular Part 1 and 2, which are served as background data. A T matrix of rank 400 was estimated using all utterances from the female speakers in background data. i-vectors and corresponding covariances were computed for each of the background, training and test utterances using the estimated T-matrix. LDA was then applied to further reduce the dimension to 200. I-vectors were then radial Gaussianised followed by length normalization as described in [12]. The post-processing procedure for covariances were the same as in [10, 11], whereby the covariances were transformed by LDA, whitened, and then scaled by the norm of the corresponding i-vector. To be consistent with [5] and reduce the number of covariances stored in memory, the TM-GPLDA model with uncertainty propagation was trained using long utterances as well 'short' utterances obtained by truncating the long utterances into 20 second segments. To reduce the total number of short duration utterances, we randomly selected around 25% of short duration utterances from each long utterance.

Table 1 presents the performances when using different scaling factors. To simplify the experiments, we used equal scaling factors for both long and short utterances and a limited number of values were compared. From the table we can find that, compared with no scaling, results with scaling factors gain significant improvements for all three conditions. It is also found that when continuing to increase the scaling factors, the model tends to overfit. The scaling factors are then selected as $1/2$ for the rest of the experiments.

Table 1. *Performance (equal error rate (EER)) of TM-GPLDA with uncertainty propagation on SRE'10 8CONV-10SEC and additional 5sec and 3sec conditions (female speakers only) with different values of scaling factors with 500 speakers in training data.*

(λ_l, λ_s)	Test duration		
	10s	5s	3s
(1,1)	8.04	13.20	17.31
$(\mathbf{1/2, 1/2})$	6.60	11.75	16.10
(1/3, 1/3)	6.60	11.48	17.09
(1/6, 1/6)	6.36	12.06	17.59
(1/12, 1/12)	6.33	12.06	17.27

Table 2 summarizes performances with difference number of training speakers in training data. As in [7], it was found that TM-GPLDA works better in severely mismatched conditions like the 5 seconds and 3 seconds conditions. It is also found that when there are enough speakers (e.g. 2000), GPLDA with uncertainty propagation outperformed the baseline, the standard TM-GPLDA and TM-GPLDA with uncertainty propagation. The potential reason for this may be that uncertainty associated with i-vectors estimated from short duration utterances are much more dispersed than those from long duration utterances or insufficient short utterances were used in training the model. But when there are only a limited number of speakers (e.g. 500), the proposed method has a significant improvement over GPLDA with uncertainty propagation as well as the standard TM-GPLDA.

Table 2. *Performance (EER) of different systems on SRE'10 8CONV-10SEC and additional 5sec and 3sec conditions (female speakers only) with different numbers of training speakers.*

500 speakers			
	Test Duration		
	10s	5s	3s
GPLDA	5.67	12.56	20.87
GPLDA_UP	7.75	13.22	19.01
TM-GPLDA	8.32	12.06	20.10
TM-GPLDA_UP	6.60	11.75	16.10
1000 speakers			
GPLDA	6.79	13.07	19.10
GPLDA_UP	5.03	11.07	16.34
TM-GPLDA	6.51	13.02	16.81
TM-GPLDA_UP	6.05	11.58	16.12
2000 speakers			
GPLDA	5.76	12.08	19.10
GPLDA_UP	5.02	10.81	15.54
TM-GPLDA	5.99	11.93	16.50
TM-GPLDA_UP	6.03	11.56	16.58

5. Conclusions

In this paper, we incorporated uncertainty propagation into twin-model PLDA (TM-GPLDA), in order to take into account both, the differences between distributions from long and short utterances and the uncertainty of i-vector. Modeling, training and scoring equations have been developed and the efficacy of the proposed technique was validated on the NIST SRE'10 8CONV-10SEC task (female trials) and additional shorter duration test conditions were created using the truncated 5 and 3 seconds test data. The results show that the proposed technique is particularly advantageous when there are a limited number of speakers in the training data. Duration compensation can also be carried out in the speaker factor space in TM-GPLDA with uncertainty propagation, and will be pursued in our future work.

6. References

[1] A. Larcher, K. A. Lee, B. Ma, and H. Li, "Text-dependent speaker verification: Classifiers, databases and RSR2015," *Speech Communication,* vol. 60, pp. 56-77, 2014.

[2] T. Kinnunen and H. Li, "An overview of text-independent speaker recognition: From features to supervectors," *Speech communication,* vol. 52, pp. 12-40, 2010.

[3] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *Audio, Speech, and Language Processing, IEEE Transactions on,* vol. 19, pp. 788-798, 2011.

[4] P. Kenny, "Bayesian Speaker Verification with Heavy-Tailed Priors," in *Odyssey*, 2010, p. 14.

[5] J. Ma, V. Sethu, E. Ambikairajah, and K. A. Lee, "Duration compensation of i-vectors for short duration speaker verification," *Electronics Letters,* vol. 53, pp. 405-407, 2017.

[6] T. Hasan, R. Saeidi, J. H. Hansen, and D. A. van Leeuwen, "Duration mismatch compensation for i-vector based speaker recognition systems," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, 2013, pp. 7663-7667.

[7] J. Ma, V. Sethu, E. Ambikairajah, and K. A. Lee, "Twin Model G-PLDA for Duration Mismatch Compensation in Text-Independent Speaker Verification," *Interspeech 2016,* pp. 1853-1857, 2016.

[8] A. K. Sarkar, D. Matrouf, P.-M. Bousquet, and J.-F. Bonastre, "Study of the Effect of I-vector Modeling on Short and Mismatch Utterance Duration for Speaker Verification," in *INTERSPEECH*, 2012, pp. 2662-2665.

[9] P. Kenny, G. Boulianne, and P. Dumouchel, "Eigenvoice modeling with sparse training data," *IEEE transactions on speech and audio processing,* vol. 13, pp. 345-354, 2005.

[10] P. Kenny, T. Stafylakis, P. Ouellet, M. J. Alam, and P. Dumouchel, "PLDA for speaker verification with utterances of arbitrary duration," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, 2013, pp. 7649-7653.

[11] S. Cumani, O. Plchot, and P. Laface, "On the use of i-vector posterior distributions in probabilistic linear discriminant analysis," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP),* vol. 22, pp. 846-857, 2014.

[12] D. Garcia-Romero and C. Y. Espy-Wilson, "Analysis of i-vector Length Normalization in Speaker Recognition Systems," in *Interspeech*, 2011, pp. 249-252.

[13] P. Bromiley, "Products and convolutions of gaussian probability density functions," *Tina-Vision Memo,* vol. 3, 2003.

[14] S. J. Prince and J. H. Elder, "Probabilistic linear discriminant analysis for inferences about identity," in *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, 2007, pp. 1-8.

[15] W.-w. Lin, M.-W. Mak, and J.-T. Chien, "Fast scoring for PLDA with uncertainty propagation via i-vector grouping," *Computer Speech & Language,* 2017.

[16] A. F. Martin and C. S. Greenberg, "The NIST 2010 speaker recognition evaluation," in *Eleventh Annual Conference of the International Speech Communication Association*, 2010.

[17] T. Kinnunen and P. Rajan, "A practical, self-adaptive voice activity detector for speaker verification with noisy telephone and microphone data," in *ICASSP*, 2013, pp. 7229-7233.

[18] J. Pelecanos and S. Sridharan, "Feature warping for robust speaker verification," 2001.

Cue Equivalence in Prosodic Entrainment for Focus Detection

Martin Ho Kwan Ip, Anne Cutler

The MARCS Institute and ARC Centre of Excellence for the Dynamics of Language (CoEDL), Western Sydney University, Penrith South, NSW 2751, Australia

m.ip@westernsydney.edu.au, a.cutler@westernsydney.edu.au

Abstract

Using a phoneme detection task, the present series of experiments examines whether listeners can entrain to different combinations of prosodic cues to predict where focus will fall in an utterance. The stimuli were recorded by four female native speakers of Australian English who happened to have used different prosodic cues to produce sentences with prosodic focus: a combination of duration cues, mean and maximum F_0, F_0 range, and longer pre-target interval before the focused word onset, only mean F_0 cues, only pre-target interval, and only duration cues. Results revealed that listeners can entrain in almost every condition except for where duration was the only reliable cue. Our findings suggest that listeners are flexible in the cues they use for focus processing.

Index Terms: prosody, entrainment, focus, speech perception

1. Introduction

Humans use prosody to signal information structure, and possibly universally [1]. Speech perception involves a number of mental challenges where listeners not only need to process the segmental features that make up the words and phrases in the speech stream, but also the prosodic features that determine the wider discourse structure and the speaker's intended message. On this view, attending to prosody may be a useful strategy for finding the most important highlighted part of the utterance, and research has indeed shown that prosodically focused words are more perceptible [2], are recognised more rapidly [3], are processed more deeply in lexical activation [4], and are better retained in memory [5, 6].

However, it remains unclear whether some prosodic cues (e.g., F_0 versus duration) may prove more informative to listeners' processing of information structure. In earlier experiments [7, 8], Cutler and colleagues discovered that listeners could anticipate an upcoming accented word by entraining to various features in the utterance prosodic contour. Using a phoneme detection task, Cutler and colleagues asked participants to listen to a series of sentences and respond as fast as they could to words that began with a specified phoneme stop target (e.g., /d/ in "duck"). Listeners responded faster to the target in sentences where the preceding intonation contour predicted high stress on the target-bearing word, compared to sentences where the intonation predicted low stress. Importantly, response times were still faster for sentences with predicted high stress contexts, even when the original target words in both contexts were replaced by an acoustically identical neutral version of the same words. Since the only difference was in the preceding intonation, it was concluded that listeners can already entrain with the cues in the preceding prosody to anticipate an upcoming focus before they receive the acoustic signals of the focused word.

Subsequent experiments [9] using the same phoneme detection paradigm revealed that listeners can still forecast an upcoming focused word even when the F_0 information in the preceding prosody is rendered uninformative (by being monotonised). Similarly, listeners can still process upcoming focus when the duration of the closure before the burst of the target stop phoneme is controlled. Building on these findings, the present paper seeks to further examine the role of different prosodic information by using natural speech from sentences recorded by different speakers who happened to have used different prosodic cues in producing the same set of stimuli.

2. Experiment 1

2.1. Method

2.1.1. Participants

The sample consisted of 22 native speakers of Australian English (M_{age} = 24.23 years, SD = 8.76 years; 15 females). All of the participants reported that they were born and raised in Australia.

2.1.2. Materials

Twenty-four unrelated experimental sentences were recorded in three versions by a female native speaker (see Figure 1). In the first version, the target-bearing word received emphatic stress. In the second version, emphatic stress was instead placed on a word that occurred later in the sentence than the target-bearing word, which, as a result, received very reduced stress. In the third version, the target-bearing word and the sentence as a whole were produced in a neutral manner. In all of the experimental sentences, the phoneme target was a voiceless aspirated bilabial stop [p^h] occurring at the start of the target word's first syllable (e.g., [p^hi:nʌts] "peanut").

Using Praat [10], the target-bearing words were excised from all three versions of each experimental sentence. The high- and low-stressed target-bearing words from the first and second versions were replaced by an acoustically identical token of the same target word from the neutral version. Thereby, two experimental conditions were constructed, each containing one version of each of the 24 spliced experimental sentences, plus an additional set of 24 filler sentences. The experimental sentences with predicted high versus predicted low stress were counterbalanced across the two conditions. To avoid interference between the sentences, sentence beginnings were varied and semantic content that could be associated with another sentence in the set was avoided. In addition, apart from the target-bearing word, none of the sentences had any additional occurrence of the target phoneme or any other stop phonemes similar to the target phoneme (e.g., [b]). All of the sentences were produced at a natural fast-normal rate.

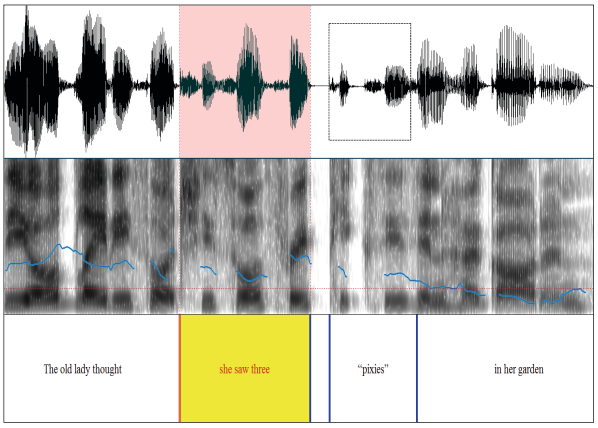
Target: [p^h]

(a) The old lady thought she saw three **[PIXIES]** in her garden.

(b) The old lady thought she saw three pixies in her **[GARDEN]**.

(c) The old lady thought she saw three pixies in her garden.

(a)



(b)

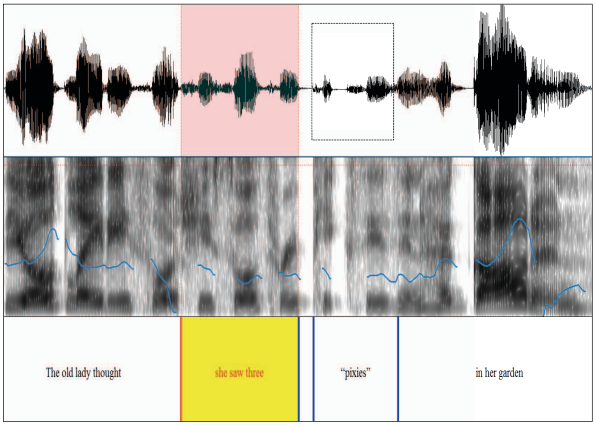


Figure 1: *Waveforms and pitch contours of an example experimental sentence in predicted high (a) and low (b) contexts; text (c) gives the neutral context. The shaded portion – three syllables preceding the target-bearing word – was analysed acoustically (section 2.1.2).*

We conducted acoustic analyses of the stimuli experimental sentences based on simultaneous inspection of the waveform and the spectrogram in Praat. Segments consisting of around three to four or five syllables before the onset of the target-bearing word were annotated and duration, mean F_0, maximum F_0, and F_0 range were measured. We also measured temporal signals such as the pre-target interval, the part of the utterance between the onset of the target-bearing word and the offset of the word before it (usually around 60 to 100 milliseconds). Results show that, for all measurements, the preceding intonation contours of sentences with predicted high stress contexts were significantly higher than the sentences with predicted low stress.

2.1.3. Procedures

Participants were tested in a sound-attenuated booth at the MARCS Institute, Western Sydney University. The phoneme-detection task was administered using E-Prime software on a laptop computer, with attached to it a set of headphones and a Chronos USB-based device for button pressing. Participants were told that the experiment aimed to examine listeners' memory and language comprehension. All participants were told that they would listen to a series of sentences and had two tasks: first, pay careful attention to the meaning of each sentence, and second, press the button as soon as they heard a word that began with the target phoneme. Participants received two practice trials and feedback before starting the actual experiment. At the end, all participants completed a follow-up recognition test in which they were asked to judge whether or not each of the 20 sentences in the list was from the experiment. We only included data from participants who scored 65 percent or above in the test.

2.2. Results and Discussion

Response times (RT) longer than 2500 milliseconds were excluded from final analyses, because such a delayed response may indicate a reprocessing of the sentence [7]. A two-tailed within-subjects t-test with an alpha threshold of .05 was conducted to assess the difference in RT between the predicted high versus low stress sentences. RTs were significantly faster in predicted high stress sentences (M = 414.92, SD = 71.68) compared to sentences with predicted low stress (M = 447.09, SD = 59.81), $t(21) = 2.83$, $p = .010$ (see Figure 2).

With respect to detection accuracy, we performed a two-tailed binomial sign test to determine whether participants were more likely to miss a button press to the phoneme target in sentences with predicted low stress than in predicted high stress. In total, there were one miss in predicted high stress contexts and five misses in low stress contexts, which was not statistically different from chance, $p = .219$ (see Table 1).

Consistent with previous studies, the results revealed that Australian English speakers can entrain with the preceding contour to forecast an upcoming focused word. However, because the acoustic analyses of the stimuli revealed significant differences for all measurements, it remains unclear as to whether some types of cues are more informative than others. Therefore, we conducted a second experiment using the same sentences produced by a different speaker.

3. Experiment 2

3.1. Method

3.1.1. Participants

We recruited a new sample of 23 native speakers of Australian English (M_{age} = 22.16 years, SD = 5.37 years; 17 females).

3.1.2. Materials and Procedures

The procedures and sentences were identical to those in the previous experiment, only this time, the sentences were recorded by another female native speaker. Acoustic analyses of the experimental sentences only revealed significantly higher mean F_0 in the predicted high stress sentences. It is important to note that no explicit instructions were given for the speaker to produce the sentences in any particular way (e.g., produce the preceding prosody with higher pitch).

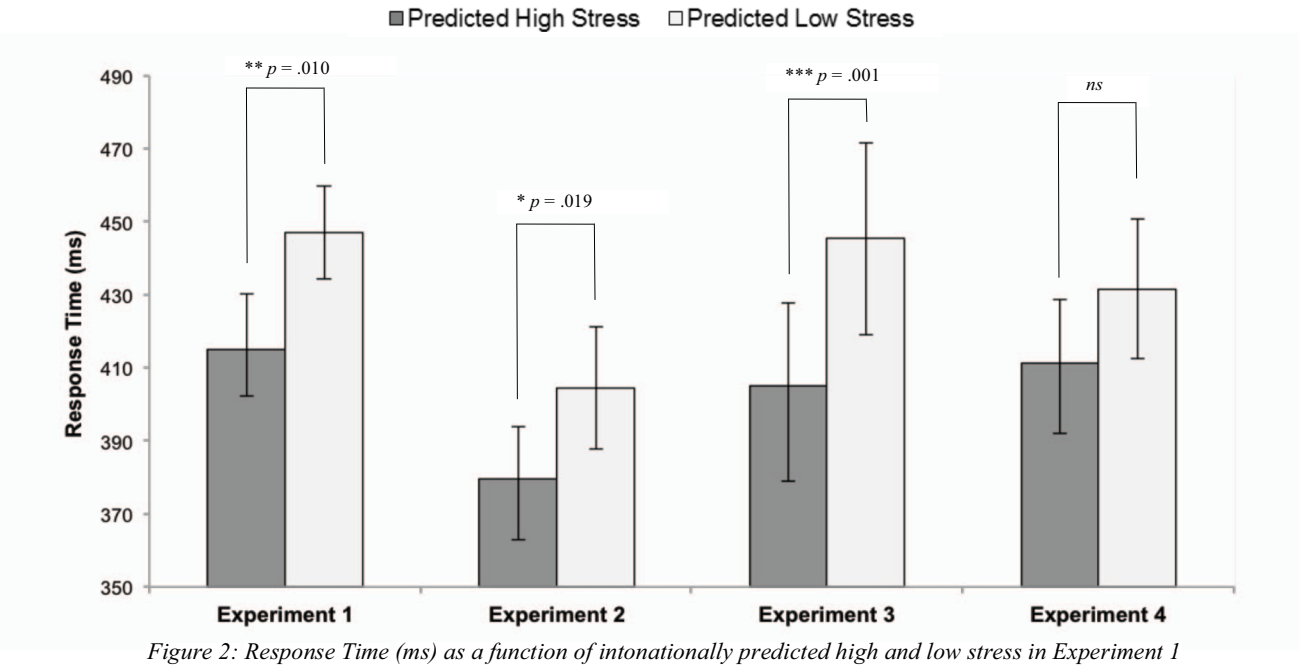


Figure 2: Response Time (ms) as a function of intonationally predicted high and low stress in Experiment 1 (with significant acoustic differences in mean F_0, maximum F_0, F_0 range, pre-target interval, and overall duration), Experiment 2 (significant acoustic difference only in mean F_0), Experiment 3 (significant difference only in the pre-target intervals), and in Experiment 4 (with only significant difference in overall duration). Error bars indicate standard error of the mean.

3.2. Results and Discussion

Consistent with the results from Experiment 1, participants' RT in Experiment 2 was faster for predicted high stress sentences (M = 379.65, SD = 68.12) compared to low stress sentences (M = 404.52, SD = 80.44), $t(22) = 2.54$, $p = .019$. In terms of accuracy, there was one miss and one false alarm (i.e. responding before the target phoneme occurred) for the predicted high stress sentences and only one false alarm for the low stress contexts.

The results indicate that listeners are as likely to use the cues from the preceding intonation regardless of whether there is a combination of many different cues (as in Experiment 1) or whether there is only one type of cue (Experiment 2). However, it is still an open question whether listeners of Australian English can still entrain if the most informative cue in the preceding prosody is not F_0-based. In the following experiments, we used the same set of sentences recorded by speakers who happened to have signalled upcoming focus using mostly duration-based cues.

4. Experiment 3

4.1. Method

4.1.1. Participants

There were 23 native speakers of Australian English (M_{age} = 22.04 years, SD = 6.80 years; 19 females).

4.1.2. Materials and Procedures

All sentences and procedures were identical to the previous experiments. Acoustic analyses of the experimental sentences recorded by the third female native speaker only revealed significantly longer pre-target intervals before the target words in the predicted high stress sentences.

4.2. Results and Discussion

Consistent with the results from the previous experiments, RT was faster for predicted high stress sentences (M = 405.19, SD = 108.50) compared to low stress sentences (M = 445.37, SD = 126.41), $t(22) = 3.96$, $p = .001$. In terms of accuracy, there were only two misses and one false alarm for the predicted low stress sentences.

5. Experiment 4

5.1. Method

5.1.1. Participants

These were 22 college-aged native speakers of Australian English (16 females).

5.1.2. Materials and Procedures

We used the same procedures and sentences from the previous experiments using stimuli produced by a fourth female speaker. Acoustic analyses of the experimental sentences from this speaker only revealed significant differences in duration, such the preceding part of predicted high stress sentences three to five syllables before the onset of the target-bearing word were longer (i.e. produced slower) than the preceding parts of the low stress sentences. There were no significant differences in the pre-target intervals or in any of the F_0 measures.

5.2. Results and Discussion

In striking contrast to the previous experiments, there was no significant RT difference between the predicted high versus low stress sentences, $t(21) = 0.96$, $p = .346$, although in the same direction. In terms of accuracy, both the predicted high and low stress sentences had an equal number of misses and false alarms (i.e. three misses and one false alarm).

Table 1. *Number of misses as a function of predicted high versus low stress contexts in Experiments 1 to 4.*

	Predicted High Stress	Predicted Low Stress
Experiment 1	1	5
Experiment 2	1	0
Experiment 3	0	2
Experiment 4	3	3

6. General Discussion

The present series of experiments provides a useful insight into how listeners use different prosodic information to detect an upcoming focused word. Consistent with previous findings, we demonstrate that listeners of Australian English can entrain with a variety of prosodic cues to forecast the location of an upcoming focused word in the utterance intonation contour. Results from Experiments 1 and 2 show that sentences that were recorded by the speaker who only consistently produced one type of cue (e.g., mean F_0) to distinguish predicted low and high stress contexts were just as likely to facilitate prosodic entrainment as the sentences produced by the speaker who produced a variety of cues. Further, in Experiment 3, having only pre-target interval as a significant temporal cue further supports the view that F_0 is not a necessary component of the preceding prosody for focus detection. However, Experiment 4 revealed that preceding prosody with longer duration (i.e. slower speech) before the predicted focus can be insufficient to support listeners' prosodic entrainment.

Overall, our findings indicate that although speakers can differ in their prosodic production, listeners are generally flexible in their use of the various prosodic information. Prosodic entrainment to locate focus may be justified by its value as listening strategy for everyday communication and semantic processing [11]. Irrespective of language or culture, holding a conversation presents a number of mental challenges. For one thing, conversational utterances tend to be fragmentary and elliptical [12]. At the same time, there is much uncertainty with respect to how a dialogue will unfold, and listeners often need to constantly organise and update their current discourse model. Given that accented words are generally the semantically most central part of the sentence, entraining to intonation contours to detect focus may therefore provide a headstart for listeners in navigating the utterance information structure early on, making it a strategy useful for all listeners for maintaining a socially effective conversation. On this view, prosodic entrainment could be understood as a comprehension process where listeners could attend to whatever cue they encounter in the speech stream to process the semantically highlighted part of speaker's message.

Of particular note are the results of Experiments 3 and 4, where listeners could successfully forecast an upcoming focused word when the length of the pre-target interval was informative, but not when there was a difference in overall duration of the preceding syllables. We speculate that one of the reasons for the lack of entrainment in Experiment 4 could be because the duration cues were in conflict with other prosodic information (e.g., preceding prosody having longer duration but low F_0) [13]. The pre-target intervals in Experiment 3 may be informative temporal cues because they represent an intake of breath or pausing before the focused word, which is in line with previous research showing that speakers tend to pause to single out new information [e.g., 14].

Future research can also assess whether listeners' flexibility in prosodic entrainment could also partly be based on a statistical learning mechanism. For example, one way in which listeners can use the different cues is by extracting the statistical information about the types of prosodic cues that are characteristic of a particular speaker.

7. Conclusion

Our findings provide evidence that (1) individual speakers within a given language (i.e., Australian English) can differ in the prosodic cues they display, (2) despite these differences, listeners can entrain with almost any cue or combination of cues in the speech signal to efficiently anticipate an upcoming focused word, and (3) it is unlikely that there is a hierarchy of cues in terms of how well they facilitate prosodic entrainment.

8. Acknowledgements

We acknowledge financial support from the ARC Centre of Excellence in the Dynamics of Language (CE140100041**).** We thank Mark Antoniou and Chris Carignan for technical advice. We also thank Matthew Stansfield for his support when we set up a student club to recruit research participants.

9. References

[1] Bolinger, D. L., "Intonation across languages", in J. Greenberg [Ed], Universals of Human Language II: Phonology, 471-524, Stanford University Press, 1978.

[2] Lieberman, P., "Some effects of semantic and grammatical context on the production and perception of speech", Lang. Speech., 6(3): 172-187, 1963.

[3] Cutler, A. and Foss, D. J., "On the role of sentence stress in sentence processing", Lang. Speech., 20(1): 1-10, 1977.

[4] Norris, D., Cutler, A., McQueen, J. M. and Butterfield, S., "Phonological and conceptual activation in speech comprehension," Cognit. Psych., 53(2):146-193. 2006.

[5] Fraundorf, S., Watson, D. G. and Benjamin, A. S., "Recognition memory reveals just how CONTRASTIVE contrastive accenting really is", J. Mem. Lang., 63(3): 367-386, 2010.

[6] Kember, H., Choi, J. Y. and Cutler, A., "Processing Advantages for Focused Words in Korean", Speech Prosody Proc., 702-705, 2016.

[7] Cutler, A., "Phoneme-monitoring as a function of preceding intonation contour", Percep. Psychophys., 20(1): 55-60, 1976.

[8] Akker, E. and Cutler, A. "Prosodic cues to semantic structure in native and nonnative listening", Biling: Lang. Cogn., 6(2): 81-96, 2003.

[9] Cutler, A. and Darwin, C. J., "Phoneme-monitoring and preceding prosody: Effects of stop closure duration and of fundamental frequency", Percept. Psychophys., 29(3): 217-224, 1981.

[10] Boersma, P., "Praat, a system for doing phonetics by computer", Glot Intl., 5(9/10): 1381-3439, 2002.

[11] Cutler, A. and Fodor, J., "Semantic focus and sentence comprehension", Cogn., 7(1): 49-59, 1979.

[12] Garrod S. and Pickering, M. J., "Why is conversation so easy?", Trends Cognit. Sci., 8(1): 8-11, 2004.

[13] Cutler, A. "Components of prosodic effects in speech recognition," Proc. 11th Intnl. Cong. Phon. Sci., 84-87, 1987.

[14] Gee, J. P. and Grosjean, J. "Empirical evidence for narrative structure", Cognit. Sc., 8(1): 59-85, 1984.

Conversational Style Mismatch: its Effect on the Evidential Strength of Long-term F0 in Forensic Voice Comparison

Phil Rose[1], *Cuiling Zhang*[2,3]

[1]Australian National University Emeritus Faculty, Australia
[2]School of Criminal Investigation, Southwest University of Political Science & Law, China
[3]Chongqing Institutes of Higher Education Key Forensic Science Laboratory, China

philjohn.rose@gmail.com, cuilingzhang@hotmail.com

Abstract

We describe a speaker verification experiment to investigate the effect of mismatch in conversational style on the strength of evidence furnished by long-term fundamental frequency in forensic voice comparison. Non-contemporaneous recordings of informal conversations, simulated police interrogations, and information exchanges from 90 male Chinese speakers were compared within the likelihood ratio framework. Evaluation with C_{llr} and error rates shows rather poor strength and weight of evidence for matched comparisons, which degrades still further with mismatched comparisons. This suggests that long-term F0 should only be used forensically in conjunction with other features.

Index Terms: forensic voice comparison, likelihood ratio, long-term F0, Chinese, validation, weight of evidence

1. Introduction

In forensic speaker identification, the expert typically compares questioned and known voice samples to help determine whether the questioned voice has come from the known speaker. Usually the questioned sample is from an offender and the known sample from a suspect, and the beneficiary is a fact-finder (judge or jury), an investigating authority (police), or legal representative [1]. Although not included in the most recent general report on the scientific validity of feature comparison methods in forensic science [2], technical forensic speaker identification also typically relies on such *feature comparison* methods [3], whether the features be the mel-frequency cepstral coefficients of automatic approaches [4], or more transparent, but less powerful, acoustic-phonetic properties like formant frequencies in so-called forensic-semi-automatic speaker recognition [5].

Features used to help identify voices forensically should ideally be common and easy to extract, and relatively immune to channel distortion. But most importantly, of course, the feature must be demonstrably effective in discriminating same-speaker speech samples from different-speaker speech samples, and have been shown to do so under the conditions of the forensic case in which they are being used:

> Without actual *empirical* evidence of the ability of a forensic feature-comparison method to produce conclusions at a level of accuracy appropriate to its intended use under circumstances reasonably related to this use, an examiner's conclusion that two samples are likely to have come from the same source is *completely meaningless*." [6, pp. 1-2].

One popular acoustic-phonetic feature in forensic semi-automatic speaker recognition – its mean and standard deviation values reportedly used by 94% and 72% of forensic voice comparison experts world-wide [7] – is fundamental frequency (F0), the acoustic reflex of the rate of vibration of the vocal cords. This is because of promising results in early speaker recognition research; and also because F0 is (relatively) easily measurable and there is usually lots of it. It is relatively immune to channel distortion (for although the fundamental and H_2 might be attenuated by phone transmission, many harmonics remain in higher frequency ranges to permit estimation of the so-called missing F0. The use of intonational F0 in a real case, and equally importantly the validity of the method, is documented in [5].

As well as the many linguistic uses of F0, which encodes tone, intonation and stress, many non-linguistic factors, like state of health, are also known to affect it. This multiplicity of factors has an adverse effect on its between- to within-speaker variance ratio by increasing the latter. Since the inherent strength of forensic speaker recognition features relies primarily on their ratio of within- to between-speaker variance, one would not expect particularly good strength of evidence (SoE) from global F0 properties, and this has been demonstrated in several studies, e.g. [1,8,9]. These studies have, however, also used arguably ecologically less than valid material – or at least material less likely to occur in real cases. For example, contemporaneous recordings were used in [8], and monologs of varying, but atypically long, duration in [1, 9]. Such conditions also have the potential to overestimate the SoE of uncontrolled global F0. Finally, one relationship between suspect and offender recordings which is commonly found in real-world case-work, but which does not seem to have been tested, is mismatch in formality between the conditions under which the suspect and offender voice recordings are obtained. Typically, the suspect's recordings are taken from a formal police interview, whereas the offender's recordings are from informal conversational exchanges. This paper's aim is to test, within a likelihood ratio framework, how well F0 from natural speech performs, in particular under these mismatched conditions.

The likelihood ratio (LR) framework [10] was recently endorsed as best practice in forensic automatic and semi-automatic speaker recognition by the *Board of the European Network of Forensic Science Institutes*, representing 58 laboratories in 33 countries [11]. As far as this paper's aim is concerned, the LR framework has two merits. Demonstrating that a forensic speaker identification method actually works is called *validation*. The LR framework allows a system to be validated in a forensically realistic manner [12], and the discriminability of forensic speaker recognition systems has in fact been tested with it now for nearly two decades. Secondly, a likelihood ratio also quantifies the SoE of a particular feature or system, and from this, as will be shown, an estimate of the weight of evidence, and expected weight of evidence, can also

easily be derived [13].

2. Procedure

2.1. Database

We used a database of 90 male Chinese speakers recorded in 2011 for the purpose of aiding forensic voice comparison research and practice. The speakers were recruited from the *Chinese Police College of Criminal Investigation* 中国刑事警察学院 in Shenyang and all spoke varieties of North-Eastern Mandarin. The database was structured according to the protocol in [14], which was designed to elicit realistic (i.e. non-contemporaneous, natural speech) recordings for testing forensic voice comparison. Speakers were recorded on two occasions separated by about a month. Three different conversational tasks were recorded in sound-proofed rooms at 44.1 kHz and 16 bit resolution: a conversation, a simulated police interview, and a co-operative exercise where both speakers were given different copies of a badly transmitted fax and had to work out its contents. For the conversation and the fax tasks, speakers were paired, and communicated on a landline phone while being recorded on separate channels using lapel mikes. For the interview, each speaker was interviewed separately by a research assistant with internship experience in interrogating suspects, and was again recorded on a lapel mike. Thus there were six recordings per speaker, labeled C1 (for *conversation recording* 1), I1 (*interview recording* 1), F1 (*fax recording 1*), and C2, I2, F2.

2.2. Front-end

For convenience in extracting F0, the quiescent portions of each speaker's recordings were removed, and the remaining non-silent portions saved as separate short .wav files using the 'sound file cutter-upper' *Matlab* code [15]. A *Praat* script was written to cycle through and inspect each putative vocalization and downsample it to 8k (to eliminate fricative and aspiration noise which can affect automatic F0 extraction). The inspection was used to estimate appropriate settings for the F0 extraction (which were between 30 Hz and 300 Hz) and a voicing threshold (0.3). *Praat* extracted F0 with these settings every centisecond, using autocorrelation.

Not all non-silent portions of the recordings were deliberate utterances, of course. Visual inspection also showed that some of the short .wav files contained extraneous noise with artefactually extracted F0, or brief hesitation phonation with badly extracted F0, and these were excluded by rejecting any short .wav file with a duration less than 35 csec. In Chinese speech, the shortest utterances (usually a monosyllabic CV or CVC word) will probably be slightly less than this; but it is obviously more important to exclude F0 measurements from non-speech than include all speech. It was also noted that longer stretches of speech sometimes contained hesitation phenomena a little before or after a longer utterance, with badly extracted F0 between the utterance and the hesitations. Such examples were excluded by ensuring that the extracted F0 consisted of at least 60% of the overall duration of the portion.

R code was then written to retain only extracted F0 values, and combine the remaining portions into a single file from which a density could be estimated. Speakers differed slightly in the amount they spoke during their tasks. Consequently, differing amounts of voiced speech were obtained, ranging from ca. 60 to 630 seconds, with a mean amount of net voiced speech per speaker of ca. 220 seconds. The different tasks also resulted in different amounts of net voiced speech, with the interview having the most (mean = ca. 280 sec.) and the fax the least (mean = ca. 150 sec.). A small difference also obtained between the two non-contemporaneous recordings, with the second showing a slightly narrower range.

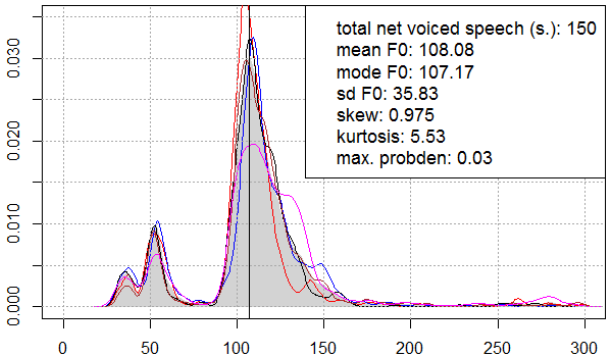


Figure 1: *Empirical kernel density distribution of long-term F0 from speaker 35's C1 recording. Coloured lines = 30 sec. replicates. X-axis = F0 (Hz), y-axis = probability density. Vertical line = modal F0.*

Each speaker's F0 data for each recording was then divided into 30-second chunks to act as replicates from which within-speaker variance could be estimated – an important part of the procedure for estimating multivariate likelihood ratios. Figure 1 shows data thus prepared for a single speaker. The grey distribution represents a kernel density for the 15000 F0 measurements from his 150 sec. voiced speech. The coloured lines show the distributions of the five different 30 sec. replicates from this. The majority of the speaker's F0 values are concentrated in a typically mildly positively skewed distribution with a modal value of 107 Hz. The speaker also has quite a lot of much lower F0 values centered at about 50 Hz which come from creaky phonation. Quite a lot of speakers show this bimodality.

2.3. Parameterisation

Previous LR-based discriminations using LTF0 distributions [1,9] have used six parameters from the F0 distribution: mean, mode, standard distribution, skew, kurtosis and maximum probability density. These were separately estimated from each speaker's 30-second replicates. One of the advantages of the MVLR formula used in this analysis is to be able to take the inevitable correlations between such parameters into account, and a principle components decorrelation, which will result in slightly degraded performance, was considered otiose.

2.4. Back-end

The multivariate kernel-density (MVKD) likelihood ratio formula [16] was used to compare the parameterized acoustics from each speaker's first recording with their second recording, and with the parameters of the other speakers' first recordings to get scores for same-speaker and different-speaker comparisons. (Only one set of different-speaker comparisons was used: between a first speaker's first recording and a second speaker's second.) The scores were then converted to likelihood ratios with logistic regression calibration using the *Focal* tool-kit [18]. A leave-one-out cross-validation was used as a strategy against overfitting,

whereby the test data were removed from the reference data for the estimation of the covariance matrices. A leave-one-out cross-validation was also used in the estimation of the logistic-regression coefficients for calibration.

The performance of an LR-based detection system like this, equivalently its validity, is currently assessed by the information-theoretic log likelihood ratio cost C_{llr} [19]. C_{llr} relates to the average amount of information the system provides to its end-user. Positive C_{llr} values below unity – the smaller the C_{llr} the better – indicate that the system has the capability of reducing the user's uncertainty in the hypothesis. Error rates for same- and different-speaker comparisons were also calculated.

With three different tasks, six different comparisons are possible: three with matched tasks, where speakers are compared using data from the same task, e.g. fax - fax; and three mismatched tasks, with comparisons based on speech from different tasks, e.g. fax - interview. For each of these mismatched tasks, two comparisons were possible from reversing the task (i.e. fax – interview and interview – fax). Nine comparisons were thus made. We are above all interested in seeing how well F0 performs with the realistically constituted data in the mismatched condition when conversation speech is compared with interview speech, which is the common type of comparison between offender and suspect in forensic reality.

3. Results

Table 1 gives the results – C_{llr}s and error-rates – for the three matched and three unmatched comparisons. The *expected weights of evidence* in the right column are explained in section 4. It can be seen first of all that, with the exception of the fax-fax comparison, all C_{llr} values are high, ranging from 0.75 to 0.87. This indicates that the long-term F0 is on average providing some information to reduce the user's uncertainty, but not much. The better performance of the fax-fax comparison (C_{llr} = 0.53) might be due to factors contributing to smaller within-speaker variation like more constrained subject matter, or less emotion. Finally, it can be seen that mismatched conditions do result in less information, with the forensically most relevant comparison, between conversation and interview, having one of the two worst performances (C_{llr} = 0.87).

Table 1. *Results. ER = error rate (%), SS/DS = same-/different-speaker comparison, EWoE= expected weight of evidence (decibans), conv = conversation, int = interview.*

comparison	C_{llr}	ER_{SS}	ER_{DS}	EWoE
matched				
conv-conv	0.79	16.7	36.5	2.0
int-int	0.75	15.6	33.5	2.4
fax-fax	0.53	11.1	23.5	4.2
mismatched				
conv-fax	0.87	20	43.5	1.2
fax-conv	0.80	14.4	37.7	2.3
conv-int	**0.87**	**18.9**	**42.1**	**1.4**
int-conv	0.85	20	40.5	1.4
int-fax	0.86	21.1	42.0	1.2
fax-int	0.85	21.1	41.7	1.3

Figure 2 shows the Tippett plot for the forensically relevant conversation-interview comparison. Same-speaker LRs increase towards the right; different-speaker LRs towards the left. The slight rightwards displacement of the equal error-rate point, which of course results in exaggeration of the different-speaker error-rate and attenuation of same-speaker error-rate, is typical but mysterious and may be related to the logistic-regressive nature of the calibration.

It can be appreciated that figure 2 depicts a feature with a very poor strength of evidence: the maximum $\log_{10}$LR observed for same-speaker comparisons was 0.46 – a LR of ca. 3 – and the mean same-speaker LR was 1.7. Even under the most advantageous conditions for the prosecution, i.e. with evidence of maximum strength and suspect just one of two who could have said the incriminating speech, the posterior probability would be just ca. 75% in favour of the suspect being the unknown speaker. Given less favourable priors – say suspect one of five possible perpetrators – and the average LR for the comparison, the posterior would be [1.7 /(1.7+4) =] ca. 30%, suggesting rational belief in a different-speaker hypothesis.

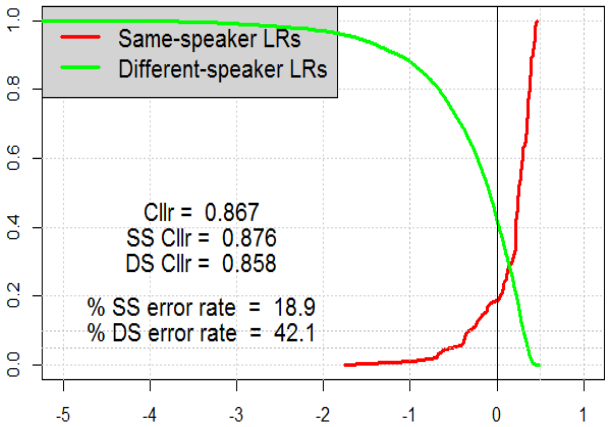


Figure 2: *Tippett plot for 90 speakers' conversation-interview comparison. X-axis = $\log_{10}$LR greater than ..., y-axis = cumulative proportion of different-speaker comparisons ~ 1-cum.prop. of same-speaker comparisons.*

4. Strength and Weight of Evidence

In English, the two properties commonly metaphorically predicated of evidence – *strength* and *weight* – both highlight the important fact that evidence can have different, continuously varying values (another metaphor!). The strength of evidence E in favour of a proposition H like *these two speech samples have come from the same speaker* is properly measured by the likelihood ratio: the ratio of the probability of getting E assuming H is true to the probability of getting E assuming it is not: $P(E|H)/P(E|{\sim}H)$ [10, pp.95ff.]. The odds form of Bayes' theorem at (1) makes it clear that the strength of evidence is also the Bayes' Factor, by which the prior odds in favour of the hypothesis [$O(H)$] have to be multiplied to get the posterior odds in favour of the hypothesis once the evidence is adduced [$O(H|E)$] [13, p.88].

$$\frac{O(H|E)}{O(H)} = \frac{P(E|H)}{P(E|{\sim}H)} \qquad (1)$$

The additive property of evidence suggested by *weight* can be modeled by simply taking the logarithm of the Bayes' Factor [13, p.89]. Turing suggested a base of 10, calling the unit a *ban,* with a *deciban* being the smallest change in weight of evidence we can conceptually process [13, p.90; 20, p.92]. For example, the maximum same-speaker Bayes Factor of ca. 3 for the conversation-interview comparison in figure 2 has a weight of about [$10*\log_{10}3$ =] 4.7 *db* in favour of it being the

same speaker. Given flat priors (i.e. 0 *db*), a LR of 3 from a comparison of LTF0 from suspect's conversations and offender's Police interview should rationally tip the scales a little in favour of your belief that the same speaker was involved. But the average LR has a hardly noticeable weight of only ($10*\log_{10}1.7$ =) 2.3 db: you would hardly notice the scale moving!

As with C_{llr}, an *overall* evaluation of weight is perhaps more informative than evaluation for individual Bayes' Factors. This can be captured with an *expected weight of evidence* in favour of the same-speaker hypothesis ($EWoE_{ss}$). This can be conventionally estimated as at (2) from the product of the probability of the two mutually exclusive outcomes and their respective weights [13, p.91].

$$EWoE_{ss} = P(\mathrm{LR} > 1|H_{ss}) * 10 * log_{10}\left[\frac{P(LR > 1|H_{ss})}{P(LR > 1|H_{ds})}\right] - P(\mathrm{LR} < 1|H_{ss}) * 10 * log_{10}\left[\frac{P(LR < 1|H_{ss})}{P(LR < 1|H_{ds})}\right] \quad (2)$$

The expected weights of evidence from the various comparisons were given in table 1. With the exception of the fax-fax comparison they are decidedly light.

5. Summary and conclusion

This paper has shown that strength of forensic speaker recognition evidence from LTF0, already weak when conditions are matched, becomes even weaker under mismatched conditions. The expected weight of evidence of systems seeking to extract forensically useful information from comparisons of LTF0 was shown to be very light indeed. LTF0's demonstrated weakness as an acoustic phonetic parameter – at least when quantified as in this paper – does not appear to warrant its popularity, at least when considered on its own.

The poor strength of evidence from LTF0 – and it should be remembered that the parameters were obtained from quite a lot of speech – indicates that it can at the moment at best only be used in conjunction with other features. Perhaps its strength can be improved by better modeling of the LTF0 distributions. One might for example separate the creaky values from the modal, and model the latter with an appropriate non-normal (log-normal? Gamma?) distribution; certainly a means should at least be found for compensating for the mismatch.

6. Acknowledgements

This paper was conceived in 2018 when the first author was visiting professor at the *School of Criminal Investigation* at the *Southwest University of Political Science and Law* in Chongqing, China. The visit was funded from a grant under the *Chongqing Municipality Attracting Overseas Expertise Scheme* 巴渝海外引智计划, for which we would like to express our gratitude. The research was also supported by the *National Social Science Foundation of China Key Program* (Grant No. 16AYY015), *Southwest University of Political Science and Law Research Funding* (2015-XZRCXM003), and *Chongqing Social Enterprise and People's Livelihood Guarantee Scientific and Technological Innovation Special Research and Development Key Project* (cstc2017shms-zdyfX0060).

7. References

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

[2] 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, 2017.

[3] González-Rodriguez, J., Rose, P., Ramos, D., Torre, D. and Ortega-García, J., "Emulating DNA: Rigorous Quantification of Evidential Weight in Transparent and Testable Forensic Speaker Recognition", IEEE Trans. on Audio Speech and Language Proc., 15(7):2104-2115, 2007.

[4] Enzinger, E., Morrison, G.S. and Ochoa, F., "A demonstration of the application of the new paradigm for the evaluation of forensic evidence under conditions reflecting those of a real forensic-voice-comparison case", Science and Justice 56:42-57, 2016.

[5] Rose, P., "Where the science ends and the law begins – likelihood ratio-based forensic voice comparison in a $150 million telephone fraud", Int'l J. Speech Language and the Law 20(2):277-324, 2013.

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

[7] Gold, E. and French, J. P., "International practices in forensic speaker comparison", Int'l J. of Speech, Language and the Law 18(2):293-307, 2011.

[8] da Silva, R.R., da Costa, J.P.C.L., Miranda, R. K. and Del Galdo, G., "Applying base value of fundamental frequency via the multivariate Kernel-Density in Forensic Speaker Comparison", IEEE 10th Int'l conf. on Signal Processing and Communication Systems, 2016.

[9] Kinoshita, Y., Ishihara, S. and Rose, P**.,** "Exploring the Discriminatory Potential of F0 Distribution Parameters in Traditional Forensic Speaker Recognition", Intl. J. of Speech Language and the Law, 16(1): 91-111, 2009.

[10] Aitken, C.G.G. and Taroni, F., Statistics and the Evaluation of Evidence for Forensic Scientists, Wiley, 2004.

[11] Drygajlo, A., Jessen, M., Gfroerer, S., Wagner, I., Vermeulen, J. and Niemi, T., Methodological Guidelines for Best Practice in Forensic Semiautomatic and Automatic Speaker Recognition, Verlag für Polizeiwissenschaft, 2016.

[12] Ramos, D. and Gonzalez-Rodriguez, J., "Reliable Support: Measuring Calibration of Likelihood Ratios", Forensic Science International 230(1-3):156-169, 2013.

[13] Good, I.J., "Weight of Evidence and the Bayesian Likelihood Ratio", in C.G.G. Aitken and D.A. Stoney [Eds], The Use of Statistics in Forensic Science, 85-106, Ellis Horwood, 1991.

[14] Morrison, G. S., Rose, P. and Zhang, C. **"**Protocol for the collection of databases of recordings for forensic-voice-comparison research and practice", Australian J. of Forensic Sci., 44(2):155-167, 2012.

[15] Morrison, G.S., Sound file cutter-upper matlab code, http://geoff-morrison.net/documents/Sound%20File%20Cutter%20Upper%20documentation.pdf

[16] Aitken, C.G.G. and D. Lucy, D., "Evaluation of trace evidence in the form of multivariate data", Applied Statistics 53(4):109-122, 2004.

[17] Morrison, G.S., 2012. "Tutorial on logistic regression calibration and fusion: converting a score to a likelihood ratio", Australian J. Forensic Sci., 1-25, 2012.

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

[19] 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.

[20] Jaynes, E.T., Probability Theory - The Logic of Science, Cambridge University Press, 2003.

Pathologic Speech and Automatic Analysis for Healthcare Applications (Batteries Not Included?)

Brian Stasak[1], Julien Epps[1] and Aaron Lawson[2]

[1]School of Elec. Eng. & Telecom., UNSW Sydney, NSW - Australia
[2]SRI International, Menlo Park, CA - United States of America
b.stasak@unsw.edu.au, j.epps@unsw.edu.au, aaron@speech.sri.com

Abstract

As the use of machine-learning techniques expand throughout the healthcare industry, the future of automatic speech analysis holds substantial promise for a non-invasive, investigative diagnosis or monitoring method for numerous medical conditions. However, depending on the disease/disorder, there is still significant uncertainty about which elicitation speech mode or modes should be adopted as a test battery. For automatic speech analysis, the clinical test application of systematic elicited speech modes remains mostly unexplored, unchallenged, and/or under-researched across many medical diagnoses. Herein we review prevalent speech modes from the perspective of controlled speech elicitation within the imminent automatic processing healthcare domain. We identify and address the shortage of comparative data in terms of speech modes for automatic diagnostic analysis.

Index Terms: Corpus specification management; database techniques; elicitation speech modes; voice language disorders

1. Introduction

Healthcare clinicians perform speech-language evaluations to assess, diagnose, and monitor many prevalent illnesses [1, 2, 3]. Clinicians pay close attention to any speech production disruptions in articulation, phonation, voicing, resonance, and prosody, which may indicate a disorder or disease. They also have an equal interest in evaluating verbal language abilities because abnormal cognitive-linguistic behaviors may denote health concerns. Contrasting between observations of speech and language ability helps to practically localize and discern different neurological or disease-related pathologies.

Clinicians rely on a battery of evidence-based tests and their subjective assessment experience to help diagnose patients. Depending on the clinician's expertise and the patient's issue, different speech modes may be used to elicit and test against healthy speaker norms. In [4], it was suggested that speech mode test material include at least four criteria: (1) speech sounds representative of those found in spontaneous conversations; (2) considerations of the economy of speech production; (3) recording methods similar in quality across all speakers; and (4) tasks that include meaningless syllable combinations, meaningful words given out of context, or meaningful phrases with contextual relations among words.

Despite the groundwork set by [4] and others [1, 5], still today the majority of clinical speech studies contain dissimilar data collected differently in terms of number of speakers, demographics, languages, speech modes, stimuli, cognitive load, recording methods, and number of sessions or durations. Further, with so much variation in speech parameters and test materials, it is hard to establish the optimal speech mode for analysis because it is exceedingly rare that two clinical speech databases are replicates of each other.

Currently, there is neither a principal guide nor set of recommended test batteries including speech modes that adequately detail what methods are most effective for reliably assessing disorders/diseases using automatic speech analysis. Automatic speech processing researchers have advocated more transparent collection methods for clinical speech data [6, 7]. However, they have yet to propose audio collection standards with clear terminology, specific test criteria, and elicitation protocol materials based on solid clinically motivated investigations.

A major difficulty is that similar abnormal speech-language symptoms are shared and observed across many different disorders/diseases. Further, many individuals have more than one medical condition present. For instance, 25% of adults have multiple chronic conditions, and this rate rapidly increases to 75% after the age of 65 years [8]. Therefore, elicited speech modes for automatic analysis require more specific disorder/disease-dependent validation. Broader investigation of speech modes will help to: identify illness-specific speech characterizations; reduce the diagnostic number of plausible disorders/diseases; and enable automated monitoring of an isolated disorder/disease.

This paper presents an insightful review discussing elicitation speech modes drawn from a range of clinical speech-language and automatic speech processing studies. Each speech mode is described along with advantages, limitations and healthcare-related applications. Although not entirely exhaustive, this text serves as one of the first informative frameworks for speech elicitation methods within the emerging automatic clinical speech-processing domain.

2. Speech Modes

2.1. Diadochokinetic

Diadochokinetic (DDK) speech mode is a conventional component of oral clinical neurology and speech-language pathology articulatory assessment protocols [9, 10, 11]. DDK speech consists of syllables with rapid articulatory repetitions. During this mode of speech, speakers are instructed to produce a fixed number of repetitive syllable combinations as quickly as possible without articulatory errors. DDK tasks involve different combinations of consonant-vowel sounds or vice-versa that are constructed of mono-, bi-, or tri-syllabic tokens. A common DDK speech task is */pa-ta-ka/*, which contains three different stop consonant positions.

A speaker's articulatory ability during DDK speech is a measure of his/her fine-motor control and muscular dexterity. Thus, ideally the belief is that healthy speakers are able to rapidly execute this task without error, whereas speakers with illness perform poorly in terms of speed and/or articulation

accuracy. DDK speech does not encompass speakers' maximum articulatory velocity, but rather is a contrived articulatory measure of maximally rapid modulation of antagonistic muscle groups to produce alternating motion [10]. DDK speech production is tested across a wide range of medical conditions (e.g. aphasia, dementia, depression, Parkinson's disease) [7, 11, 12, 13]. While many different types of clinicians utilize DDK, unless previously diagnosed, this speech mode is a very generalized indicator for health concerns. For example, DDK speech alone cannot provide information about a speaker's *cognitive-language* abilities or exact a precise clinical condition.

In [14, 15, 16] it was noted that the DDK speech mode is different from what occurs during normal speech; for example, DDK task speeds are relatively low relative to speeds observed in spontaneous speech. Further, due to the limited successional-temporal demand, DDK tasks induce articulatory fatigue, which is scarcely observed in natural speech [14].

Despite minimal gender-effect differences in DDK performance [17], within a healthy speaker population, DDK performance had variability among speakers in syllable rate and regularity [14, 15]. Essentially, the DDK speech mode is an indicator of which articulatory strategy is employed to maximize speech and approximate-target phonemes successfully [14]. Speakers use articulatory strategies to maximize their speech production and transition per syllable. This includes generalizing the articulatory location of the consonant-vowel targets and tongue, as well as lip closure. Noticeable differences in speaker strategies and DDK performance exist depending on the speaker-task combination. In [12], DDK compensatory strategies limited identification of individuals with early onset Parkinson's disease.

Although very constrained, DDK has remained a popular speech mode for clinical speech assessment analysis because it has a clinical origin, instructive simplicity, and requires a minimal amount of data for assessment. However, [18] has suggested that DDK has shortfalls in terms of specificity to disorders, repeatability, languages, and demographics. The inter- and intra- rater reliability was also found to be lower than should be accepted for clinical diagnostic assessment [18]. Other work has revealed that speakers exhibited a greater number of speech errors for the DDK mode than is normally expected in conversational speech, advising that the DDK task is more unnaturally difficult than natural speech tasks [19, 20].

Alternatively, DDK speech can be obtained via automatic, read, and spontaneous speech modes. For instance, the word '*buttercup*' is similar in articulatory respects to non-word DDK tasks. Assessing DDK production by natural real-word tokens was found as a more useful diagnostic gauge than standard non-word DDK stimuli with children [15, 16].

Research [13] has asserted that DDK speech tasks are immune to language-dependent properties, and may be universal in characterizing speech motor disorders. But, this universal language DDK concept has since been disputed. In [17], it was discovered that significant between-language DDK variation norms exist. Thus, there is a need to clinically validate language sensitive norms for DDK speech tasks [14, 16]. Large-scale DDK language-dependent studies will help to determine trends and its effectiveness as a speech mode for wider scope clinical assessments.

2.2. Held-Vowel

The held-vowel speech mode is effective for providing fundamental information about an individual's voice quality (e.g. voicing, hoarseness, timbre) and respiratory control. In clinical audio analyses, held-vowel tasks often simply comprise the /*a, i, u*/ corner vowels. Surprisingly, despite only consisting of a single phoneme, the held-vowel mode has demonstrated usefulness in identifying and monitoring depression [7], dysphonia [21], oral cancer [22], Parkinson's disease [23], and respiratory disorders [24].

Since the held-vowel speech mode provides trivial linguistic content, it is not very useful for assessing *cognitive-language* disorders. For held-vowel analysis it is important that speaker instructions include duration and quality criteria per vowel example. To our knowledge, there is no study across any medical condition that examines differences in automatic assessment performance based on different speech held-vowel durations. However, it is known that between-session performance may vary even with the same speaker due to everyday circumstances (e.g. cold, tiredness, stress) [7].

2.3. Automatic

Automatic speech mode is both articulatory and language driven. Additionally, it deals to a great extent with explicit knowledge and relies on information retrieval. When observing automatic speech, clinicians are primarily concerned with the articulatory accuracy, speed and appropriateness of a patient's verbal response. Examples of automatic speech include counting, birthdate, alphabet, months, repeating audible sentences, yes/no responses, image identification, word opposite, and categorization tasks [25].

Automatic speech also includes sentence-level tasks, such as rule-based unprompted dialog, prompted-question answering, device command-control, and spelling diction. Automatic tasks can combine aspects of semi-read and semi-spontaneous speech. For example, such a task may include reading aloud part of a common idiom and verbalizing its completion (e.g. "*It is raining cats {and dogs}*"). Automatic speech tasks have been used to help identify and monitor patients with aphasia [11], dementia [26], depression [7], dysphonia [21], learning disabilities [27], respiratory disorders [24], and traumatic brain injuries [28].

In [25], it was shown that automatic speech typically does not engage the language cortex; however, reiterating prose passages produces activation in both Broca's and Wernicke's language areas. Thus, if analysis of an isolated *cognitive-language* area or language skill is of particular interest, automatic speech task materials require thoughtful planning. Automatic speech tasks require careful consideration in regard to a speaker's age, native language, and cultural background.

2.4. Read

The read speech mode consists of pre-selected words, sentences, or paragraph excerpts that are read aloud by a patient. Similar to automatic speech elicitation, clinicians select certain read tasks depending on the patient's articulatory or language concern. Clinicians can easily manipulate the degree of articulatory-linguistic difficulty via the text stimuli, and consequently elicit larger amounts of the target behavior of interest [29]. Other advantages to read speech tasks include simple instruction, repeatability, isolated cognitive demand, and phonetic variability control.

Read speech tasks have been used to clinically assess patients with aphasia [11], apraxia [11], depression [7], learning disabilities [28], and Parkinson's disease [12]. Among the most popular read speech excerpts for clinical speech analysis are [30-35]. However, many of these texts are

antiquated and are unnatural in terms of modern-day speech [36]. Further, [37] found that some of these texts comprise unusual English syntax, which impact speakers' testing performance.

Perhaps the strongest argument against using many of the aforementioned texts is they were not all designed or originally intended for diagnoses of multiple types of disorders – in fact, many of these texts were only intended as subjective tests for speech intelligibility [30-33, 36]. For read speech tasks there is a need to address age, cultural and diagnostic appropriateness. Additionally, pre-screening evaluations for reading ability and eyesight should be completed to further substantiate read speech task results.

2.5. Spontaneous

The spontaneous speech mode is synonymous with spoken open-thought or conversational speech. This mode is often called 'free'; however, in a clinical assessment setting restrictions (i.e. familiarity, structure, topic) usually still apply. Spontaneous speech allows the examination of an individual's explicit and tacit knowledge in parallel with a high level of organization when compared to other speech modes. The most compelling argument for analysis of spontaneous speech is its ecological validity, which represents a person's *cognitive-language* ability and unrestricted natural social behaviors [3].

Spontaneous speech is usually collected using the following tasks: interview-type questions, telling personal stories, conversing on a particular topic, summarizing a short video, opinion/review, or the Rorschach test. For spontaneous speech interviewing, even though the same questions are repeatable across different speakers, there is no guarantee that all speakers will interact using a suitable useful response for analysis. It has been shown that interviewers' demeanor, personal bias, interviewing skill, and cultural sensitivity influence how patients verbally socialize [1, 4].

Even within the spontaneous speech mode, there exists a wide range of spoken content. It is relatively unknown how this varied content impacts behavior and automatic speech processing performance. For example, spontaneous speech constitutes numerous variables related to socialization, such as: point-of-view accounts; impersonal to personal; multidisciplinary to highly specialized focused subject matter; monologue to competitive discussion; unilateral to multilateral discourse; and question-answer to unconstrained formats. Spontaneous speech also contains a high number of natural disfluencies and auxiliary speech behaviors (e.g. sighs, laughs, pauses). During spontaneous speech, speakers often openly discuss their life, family, workplace, and medical history. Therefore, a pervasive complication in using recorded spontaneous speech is maintaining patient confidentiality.

3. Additional Considerations

During interviews, clinicians explicitly or implicitly evaluate a speaker's voice quality, which can vary remarkably from person to person. These differences in the voice can be genetic, acquired, or a temporary speech attribute (e.g. age, cold, strain). Examples of voice quality include: rate-of-speech, loudness, pitch, and timbre (e.g. breathy, creaky, hoarse, tremulous). Although voice quality has many perceptual terms, it is usually measured via a continuous human-perceptual scale based on a clinician's experience.

During interviews/tests, clinicians also often evaluate the appropriateness of a particular speech type or affect using mock context-specific scenarios. Inappropriate speech types or affect can reflect psychogenic disorders (e.g. depression, stress), neurological illnesses (e.g. schizophrenia, dementia), or congenital genetic disorders (e.g. autism, hearing loss) [11].

All speech modes can be produced using a distinctive and controlled vocal technique. Besides ordinary speech, vocal techniques include: singing, whispered, motherese, elderspeak, impersonation, and speech under mental/physical stress. Although there is little known regarding the usefulness of speech various vocal techniques for clinical assessment, it has been shown that different types can impact clinical assessment results. For example, [38] found that stuttering was reduced considerably when whispering versus normal speech. Also, during whispered speech, increased airflow expenditure was helpful in isolating respiratory issues and speech articulators [39]. Studies [40, 41] suggested that singing in neurologically impaired patients provides another means for monitoring progress in patients' speech abilities.

Paralinguistics is also an additional speech parameter [4]. Emotional analysis of speech has often included opposite situations, affective versus non-affective pictures, and stressed versus non-stressed stimuli [1]. More recently, text-based speech analysis measures related to linguistic structure and word affect have been used to automatically evaluate individuals' responses. Text analytics on session transcripts have been used to monitor depression, dementia, and stress [3]. In a clinical setting, linguistic and affect measures have been used in the design of speech task materials, allowing more control over linguistic complexity and sentiment. It was demonstrated that the degree of spoken articulation effort, linguistic complexity, and word affect impacts automatic speech-based depression classification performance [42].

Another important measure often overlooked in clinical speech databases pertains to speaker description metadata. Metadata is essential to de-generalize big populations, allowing the creation of more specific speech-language speaker models. Metadata speaker trait examples that can influence clinical speech-language analysis include a speaker's personality; parent's language; accent/dialect; prior pathologies; age; height; weight; drug use (e.g. drinking, smoking); hearing; education; and profession. In addition, metadata information regarding the interviewer can also be of interest, as it has been shown in studies this influences a speaker's social behaviors [1, 4]. Metadata adds considerable value to a speech data collection because it allows for stronger result validation of specific, isolated speaker criteria. Additionally, metadata also increases the opportunity for new discrete speech-illness correlations to be discovered.

4. Conclusion

With new advances in machine learning, a major shift towards digitally automated analysis in every area of clinical healthcare is inevitable. It is crucial that metadata considerations are taken into account when designing, choosing, and executing clinical speech elicitation materials for automatic analysis. Also, rather than reinventing the wheel, future elicitation collection methods should be more targeted to help investigate the overall effectiveness of different speech modes for specific clinical analysis. When collecting and automatically analyzing healthcare speech data, it is imperative that researchers ask, "*What are we trying to capture*? *Is this speech mode clinically validated for this particular disorder/illness and are its tasks suitable?*"

Without addressing these questions, insights and advances the healthcare utility of speech may remain elusive.

5. Acknowledgements

The work of Brian Stasak and Julien Epps was supported by ARC Discovery Project DP130101094 and ARC Linkage Project LP160101360, Data61, CSIRO.

6. References

[1] Chevrie-Muller, C., Sevestre, P., & Seguier, N., "Speech and psychopathology", Lang. and Speech, vol. 28(1), pp. 57-79, 1985.

[2] Kent, R.D., "Research on speech motor control and its disorders: a review and prospective", J. of Comm. Disord., vol. 33(5), pp. 391-427, 2000.

[3] 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.

[4] Egan, J.P., "Articulation testing methods", Laryngoscope, vol. 58, pp. 955-991, 1948.

[5] 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.

[6] Baghai-Ravary, L., & Beet, S.W., Automatic Speech Signal Analysis for Clinical Diagnosis and Assessment of Speech Disorders, Springer, New York, USA, 2013.

[7] Cummins, N., Scherer, S., Krajewski, J., Schnieder, S., Epps, J., & Quatieri, T.F., "A review of depression and suicide risk assessment using speech analysis", Speech Comm., vol. 71, pp. 10-49, 2015.

[8] Centers for Disease Control and Prevention (CDC), "Chronic Disease Prevention and Health Promotion", May 2018 via: https://www.cdc.gov/chronicdisease/about/multiple-chronic.htm

[9] Fletcher, S.G., "Time-by-count measurement of diadochokinetic syllable rate", J. of Spch. and Hear. Disord., vol. 15, pp. 763-770, 1972.

[10] 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.

[11] Duffy, J.R., Motor Speech Disorders: Substrates, Differential Diagnosis, and Management, Mosby, St. Louis, USA, 2005.

[12] Harel, B., Cannizzaro, M., Cohen, H., Reilly, N., & Snyder, P., "Acoustic characteristics of Parkinsonian speech: a potential biomarker of early disease progression and treatment", J. of Neuroling., vol. 17, pp. 439-453, 2004.

[13] Padovani, M., Gielow, I., & Behlau, M., "Phonarticulatory diadochokinesis in young and elderly individuals", Arq. Neuro-Psiquiatr., vol. 67(1), 2009.

[14] Westbury, J.R., & Dembowski, J., "Articulatory kinematics of normal diadochokinetic performance", Annual Bulletin of the Res. Inst. of Logopedics and Phonetics, vol. 27, pp. 13-36, 1993.

[15] Yaruss, J.S. & Logan, K.J., "Evaluating rate, accuracy, and fluency of young children's diadochokinetic productions: a preliminary investigation", J. Fluency Disord., vol. 27(1), pp. 65-85, 2002.

[16] Icht, M., Ben-David, "Oral-diadochokinetic rates for Hebrew-speaking school-age children: real words vs. non-words repetition", Clin. Ling. & Phon., vol. 29(2), pp. 102-114, 2014.

[17] Icht, M., & Ben-David, B., "Oral-diadochokinesis rates across languages: English and Hebrew norms", J. Comm. Disord., vol. 48, pp. 27-37, 2014.

[18] Gadesmann, M., & Miller, N., "Reliability of speech diadochokinetic test measurement", International J. of Lang. & Comm. Disord., vol. 43, pp. 41-54, 2008.

[19] Jaeger, J.J., "Not by the chair of my hinny hin hin: some general properties of slips of the tongue in young children", J. of Child Lang., vol. 19, pp. 335-366, 1992.

[20] Stemberger, J.P., "Speech errors in early child language production", J. of Mem. and Lang., vol. 28, pp. 164-188, 1989.

[21] Alsulaiman, M., "Voice pathology assessment systems for dysphonic patients: detection, classification, and speech recognition", IETE J. of Res., vol. 60(2), pp. 156-167, 2014.

[22] De Bruijn, M., ten Bosch, L., Kuik, D., Quene, H., Langendijk, J., Leemans, C., & Verdonck-de Leeuw, I., "Objective acoustic-phonetic speech analysis in patients treated for oral or oropharyngeal cancer", Folia Phoniatr Logop, vol. 61, pp.180-187, 2009.

[23] Hemmerling, D., Orozco-Arroyave, J.R., Skalski, A., & Nöth, E., "Automatic detection of Parkinson's disease based on modulated vowels", In: INTERSPEECH '16, San Francisco, USA, 2016.

[24] Shrivastava, P., Tripathi, N., & Singh, B., "Assessment of respiratory disorders using speech parameters", Intern. J. Future Rev. Comp. Sci. & Comm. Eng., vol. 4(2), pp. 461-464, 2018.

[25] Bookheimer, S.Y., Zeffiro, T.A., Blaxton, T.A., Gaillard, W., & Theodore, W.H., "Activation of language cortex with automatic speech tasks", Neurology, vol. 55(8), 2000.

[26] Jorm, A., "Controlled and automatic information processing in senile dementia: a review", Psych. Med., vol. 16, pp. 77-88, 1986.

[27] Swanberg, M., Nasreddine, Z., Mendez, M., & Cummings, J., "Speech and Language", In: C. Goetz (Ed.), Textbook of Clinical Neurology (3rd Ed.), Saunders, 2007.

[28] Gurland, G. & Marton, K., "Assessment of language disorders in school-aged children", In: C. Stein-Rubin & R. L Fabus (Ed.), A Guide to Clinical Assessment and Professional Report Writing in Speech-Language Pathology, Delmar Pub., 2012.

[29] Sidtis, D., Cameron, K., & Sidtis, J., "Dramatic effects of speech task on motor and linguistic planning in severely dysfluent Parkinsonian speech", Clin. Ling. Phon., vol. 26(8), pp. 695-711, 2012.

[30] Fairbanks, G., Voice and Articulation Drillbook (2nd Ed.), Harper & Row, New York – USA, 1960.

[31] Van Riper, C., Speech Correction (4th Ed.), Prentice Hall, Englewood Cliffs, NJ – USA, 1963.

[32] Townsend, G.F., Aesop's Fables, George Routledge & Sons, London & New York, 1868.

[33] Crystal, T.H., & House, A.S., "Segmental duration in connected speech signals: preliminary results", J. Acoustic Soc. Am., vol. 72(3), pp. 705-716, 1982.

[34] 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.

[35] 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.

[36] Reilly, J., & Fisher, J., "Sherlock Holmes and the strange case of the missing attribution: a historical note on 'the grandfather passage'", J. of Spch., Lang., and Hear. Res., vol. 55, pp. 84-88, 2012.

[37] Boaz, B., 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.

[38] Perkins, W., Rudas, J., Johnson, L., & Bell, J., "Stuttering: discoordination of phonation with articulation and respiration", J. of Speech, Lang., and Hear. Res., vol. 19, pp. 509-522, 1976.

[39] Monoson, P. & Zemlin, W.R., "Qualitative study of whisper", Folia Phoniat, vol. 36, pp. 53-65, 1984.

[40] Cohen, N., "The effect of singing instruction on the speech production of neurologically impaired persons", J. Music Therapy, vol. 29(2), pp. 87-102, 1992.

[41] Peretz, I., Gagnon, L., Heebert, S., & Macoir, J., "Singing in the brain: insights from cognitive neuropsychology", Music Perception: An Interdiscip. J., vol. 21(3), pp. 373-390, 2004.

[42] Stasak, B., Epps, J., & Goecke, R., "Elicitation design for acoustic depression classification: an investigation of articulation effort, linguistic complexity, and word affect", In: INTERSPEECH '17, Stockholm, Sweden, 2017.

Forensic Voice Comparison using Long term Fundamental Frequency in Male Australian English Speakers

Georgia Johnston [1], Shunichi Ishihara [2,3]

[1] College of Arts and Social Sciences, Australian National University, Canberra
[2] College of Asia and the Pacific, Australian National University, Canberra
[3] Speech and Language Laboratory, Australian National University, Canberra
g.anne.j@gmail.com, shinichi.ishihara@anu.edu.au

Abstract

This paper investigates the efficacy of fundamental frequency (F0) to discriminate between speakers for the purposes of likelihood ratio-based forensic voice comparison (FVC) in Australian English. The experimental results indicate that the usefulness of F0 as an FVC feature in English is comparable with that of other languages. The results also indicate that registers (e.g. formal and casual) may differently contribute to the performance of the FVC system.

Index Terms: forensic voice comparison, F0, Australian English, likelihood ratio, different registers

1. Introduction

This paper investigates to what extent long term fundamental frequency is useful to estimate likelihood ratios in male Australian English speakers. Previous studies have reported the usefulness of fundamental frequency (F0) for forensic voice comparison (FVC) in pitch-accent languages (e.g. Japanese [1]) and tonal languages (e.g. Cantonese [2]). However, the linguistic function of pitch in these previously-studied languages is quite different from English, which is a stress-accent language. As English is the national language of Australia and many other countries, it is important to test the usefulness of F0 for the purposes of FVC.

The database used in this paper is populated by more than 500 Australian English speakers, providing three different elicitation tasks [3]. Only two elicitation task samples were used to see whether there was any difference in the performance of the FVC system according to register differences. The database uses only male speakers as there has been a consistent trend of males offending at approximately three times as much as females [4].

There are pros and cons for F0 being used as an FVC feature. F0 meets three of the integral criteria advanced by Nolan [5] which are that the data is easily available, measurable and robust in terms of being unaffected by environmental effects. F0 correlates to pitch and the issue with pitch is that it can be consciously and unconsciously changed. Despite this, although F0 is not necessarily a strong feature in terms of the magnitude of the LR that can be obtained [cf. 6], F0 is still a valuable resource as FVC is not conducted with only one feature of voice in mind; the LR based on F0 can be combined with the LRs based on other acoustic features.

Variation within speakers must also be taken into consideration when designing any FVC experiment. F0 variations within speakers can be affected by emotion, such as panic or grief [7], as well as physical conditions such as inebriation.

In this study, each recording is modelled using the statistical features (mean, sd, skew, kurtosis, modal F0 and modal density) based on the distribution of the sampled F0 values. In this study, it is called the "six parameter method" [1].

2. Likelihood ratio

This study is an LR-based FVC study. In the context of forensic science, as given in (1), an LR is the probability (P) of the evidence (E) occurring if an assertion is true (e.g. the prosecution hypothesis (H_p) is true), divided by the probability that the same evidence would occur if the assertion is not true (e.g. the defence hypothesis (H_d) is true) [8].

$$LR = \frac{p(E|H_p)}{p(E|H_d)} \quad (1)$$

In FVC, the LR value would indicate the probability of viewing the difference between two speech samples (e.g. the offender and suspect speech samples) if they had come from the same speaker, relative to the probability of viewing the same evidence if the two speech samples had come from different speakers. The LRs that are higher than one (LR > 1) support the H_p, and those that are lower than one (LR < 1) support the H_d. The further away the LR is from unity (LR = 1), the stronger it supports either hypothesis.

3. Methodology

3.1. Database

The database used to conduct the experiment was compiled by Morrison [3] specifically to enable researchers/forensic scientists to perform FVC research and casework. As mentioned previously, the database consists of more than 500 Australian English speakers from the ages of 18 to 65 both male and female. They all participated in three elicitation tasks which were recorded in two or more non-contemporaneous sessions for the majority of participants.

The elicitation tasks were designed to obtain different types of speech registers, and simulate common scenarios which would likely occur in real life forensic work.

The first task, coded as CNV, was a conversational task with two participants who were known to each other were separated in different booths and used a two way intercom with a telephone style handset. They were then instructed to have a conversation about any topic they wished to. This was in an effort to obtain a conversation style as natural as possible for each participant.

The second task was an information exchange task where each participant had the corresponding missing information their interlocutor had.

Speaker 514	mean	sd	skew	kurtosis	modal f0	modal density
S1	132.41	47.24	0.96	5.01	124.25	0.02
S2	137.27	45.65	0.87	5.21	124.23	0.01

Table 1: *Six parameter values of Speaker 514 for the two sessions (S1 and S2).*

Lastly the third task, coded as INT, was a mock police interview where the researcher conducted the interview, and audio was recorded on lapel microphones. The researcher conducting the interview was under strict instructions to stop speaking as soon as the participant started, and the use of lapel microphones instead of a table microphone helped to reduce the amount of audio being affected by the researcher's voice.

Samples from the CNV and INT tasks were chosen in this study due to the nature of the tasks being close to what a forensic linguist may actually deal with for a case and being different in their registers.

In order to see how the amount of samples influences the performance of the FVC system, only the speakers whose recordings were longer than 120 sec voiced segments in two contemporaneous sessions have been selected, resulting in 100 speakers. The average duration of speech for obtaining 120 sec voiced segments is 3.38 minutes.

The 100 speakers were separated into three mutually-exclusive databases of test (34 speakers), background (33) and development (33) databases. The test database is used to simulate the comparison of offender and suspect samples. Out of the 34 speakers of the test database, 34 same-speaker (SS) and 1122 different-speaker (DS) comparisons are possible. The background database is used to generate statistical information for typicality whilst the development database is used for estimating the calibration weights.

3.2. Fundamental Frequency Extraction

Using the ESPS tool of the Snack Sound Toolkit, F0 was extracted from each recording, which was sampled and downloaded at 16kHz, at every 0.02 sec with a hamming window, shifting every 0.01 sec.

3.3. Six Parameter Method

All sampled F0 values are pooled together for each recording to obtain the six parameters of mean, sd, skew, kurtosis, modal F0 and modal density for modelling the recording [1]. Figure 1, which includes the long term F0 distributions of the two non-contemporaneous sessions of Speaker 514, shows a high degree of similarity in the distributional shapes between the two sessions.

Table 1 contains the calculated six parameter values of the long term F0 distributions given in Figure 1. Reflecting the similarity between the two distributions, the six parameter values of Table 1 are also similar between the two sessions.

The six parameters were calculated using a different amount of F0 samples: 5, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 110 and 120 sec of voiced segmental duration. In the current studies, linear and $\log_{10}$ F0 are used.

3.4. Likelihood Ratio Calculation

It is common knowledge that F0 is not normally distributed. In addition, some of the variables, or parameters, we are using in this study are correlated, which makes estimating an LR value quite complex. The Multivariate Kernel Density (MVKD) formula posited by Lucy and Aitken [9] allows the researcher to compute a single LR value from multiple variables, whilst accounting for correlation. Morrison [10] supports the usage of MVKD in FVC stating that it should be considered standard procedure in the calculation of LRs. In this study, the MVKD formula is used to estimate LRs.

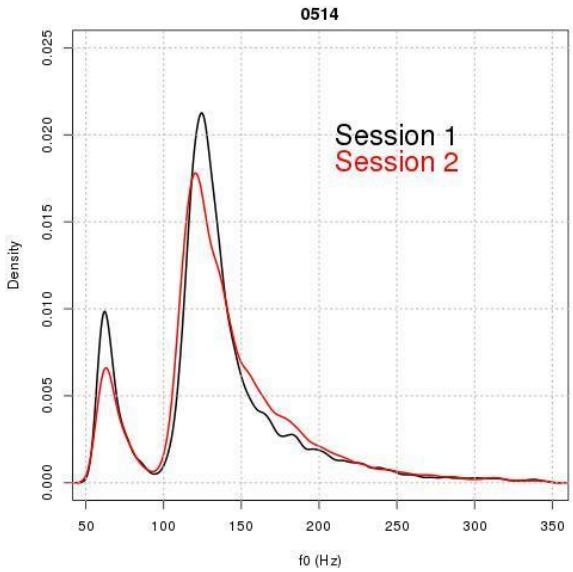


Figure 1: *Example long term F0 distributions of two non-contemporaneous sessions from Speaker 514*

3.5. Metrics of Assessment

To assess the outcomes of the experiment, we have used three measures. First is the Equal Error Rate (EER) which is a measure of discriminability in an LR system, it is the point where the SS comparison and DS comparison error rate is the same. The second assessment is the log likelihood ratio cost (C_{llr}), which is a cost-based function in which the contrary-to-fact LRs are penalised heavily; the closer the C_{llr} is to zero the better the system is performing in terms of validity. The final metric is the 95% Confidence Interval (95%CI) which is a measure of reliability of a system and also provides a metric for how precise a system is. The lower the 95%CI value, the less distributed the performance.

4. Results

The C_{llr} values of the CNV and INT are plotted in Figure 2 as a function of the voiced duration of 5, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 110 and 120 sec. The results based on the $\log_{10}$ F0 are given in this section as the results based on the linear F0 showed an essentially identical trend.

Except the C_{llr} values of the INT with 40 and 120 sec of voiced duration, although there are some minor ups and downs in the C_{llr} values, the validity of the system shows an improving trend as a function of voiced duration, with a large improvement until 40-50 sec of voiced duration, after which the improvement is minimal. We double-checked the samples of 40 and 120 sec for the INT, but could not find any

irregularities. This may be due to the instability of the MVKD formula [11] and not having enough erroneous classifications for the SS and DS comparisons for C_{llr}, but it requires further investigation.

There seems to be a difference in performance between the two elicitation tasks: CNV and INT. It is evident from Figure 2 that the CNV performs better than the INT in most cases.

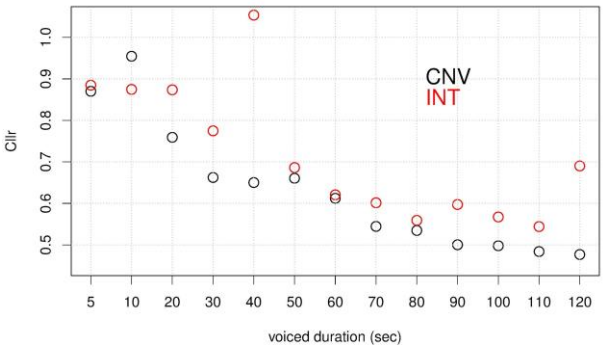


Figure 2: *C_{llr} values are plotted as a function of voiced duration, separately for CNV and INT.*

The 95%CI values of the CNV and INT are plotted as a function of the different voiced durations in Figure 3. Again the 95%CI values of the INT with 40 and 120 sec of voiced duration show non-conforming behaviours. However, it can be generally observed from Figure 3 that 1) the reliability tends to deteriorate as a function of the voiced duration up to ca. 50 sec, and 2) after ca. 50 sec, they are relatively stable or show a downward trend, in particular for the CNV. In terms of 95%CI too, the CNV performs overall better than the INT.

Some studies [12] reported that 95%CI is negatively correlated to C_{llr}. However, the results of the current study are intriguing in that 1) the CNV performs overall better than the INT for both the validity (C_{llr}) and precision (95%CI), and 2) the C_{llr} continues to improve as a function of voiced duration (with lesser degree after 40-50 sec), while there seems to be a ceiling around 50 sec in terms of 95%CI, after which even a moderate downward trend can be observed for the CNV.

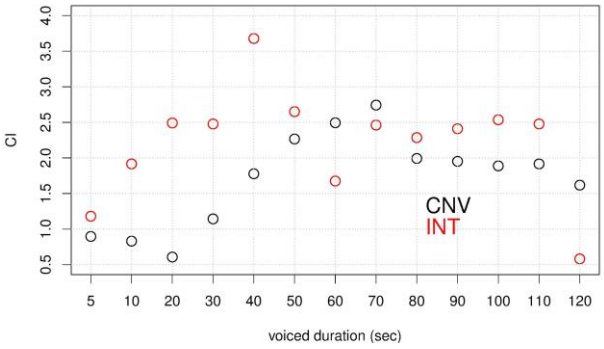


Figure 3: *95% CI values are plotted as a function of voiced duration, separately for CNV and INT.*

The above observations about the 95%CI values indicate that the validity and precision of an FVC system may tend to be negatively correlated up to a certain amount of data, but after that, both of them can improve as a function of the amount of data. Needless to say, this point warrants further investigation with more experiments with other datasets.

Another helpful way to visualize the results is by using a tippet plot to see the magnitude of the derived LRs. Figure 4 contains Tippett plots of the CNV with 5 and 120 sec.

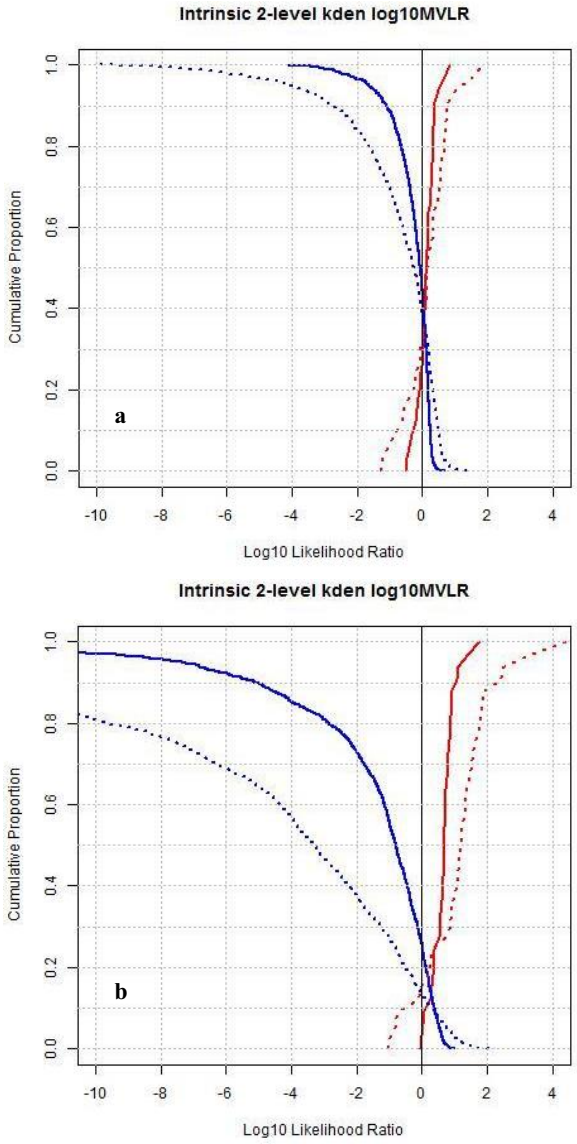


Figure 4: *Tippet plots showing the magnitude of LRs for CNV with 5 (panel a) and 120 (panel b) sec of voiced duration. Solid curves = calibrated LRs; Broken curves = uncalibrated LRs; Red = SSLRs; Blue = DSLRs.*

In Figure 4, the broken curves represent the uncalibrated LRs and the solid curves represent calibrated LRs. The red curves represent SSLRs and the blue curves DSLRs.

It can be observed from Figure 4 that the LRs are very well calibrated even before any calibration is applied. The impact of calibration is mainly for scaling in that the magnitude of the calibrated LRs (solid curves) is more conservative than that of the uncalibrated LRs (broken curves). The impact of the increase in the amount of data (i.e. voiced duration) is clear from Figure 4 in that 1) the system discriminability is substantially better with 120 sec of voiced duration (EER = 0.14) than with 5 sec of voiced duration

(EER = 0.36). Moreover, the magnitude of the consistent-with-fact LRs is (far) greater with 120 sec of voiced duration than with 5 sec of voiced duration, in particular for the DSLRs.

5. Discussions and Future Studies

One of the most interesting results that was revealed through the experiments is how the speech register affected the performance of the system; the CNV performed better in both validity and reliability than the INT.

One potential factor could be that relaxed conversation between two speakers known to each other may allow for more expressive communication, providing a more individuating F0 distribution. Gibbon [13] supports this theory and posited that intonation in language is linked to semantic structures, and that the notion of emphasis and contrast in 'tone' is a result of accentuated elements that are specific to semantic relations in the discourse. This infers that the topic of conversation greatly influences the intonation/F0 contours in the speech sample. This result may seem counterintuitive as there is more opportunity for a wider range in pitch for the CNV would lead to more within speaker variation and thus more erroneous SS comparisons but this hasn't occurred in this study. This avenue of investigation certainly warrants further research.

It is useful to compare the results of the current study to those of other similar studies. Table 2 is a concise summary of the results of the current and previous studies.

Papers	C_{llr}	EER
Li & Rose 2012 [2]	0.86	40%
Zheng & Rose 2012 [14]	0.52	ca. 22%*
Kinoshita et. al 2009 [1]	N/A	ca. 17.5%*
CNV with 40 sec duration	0.65	24.3%

Table 2: *A summary of the current and previous studies for comparison. *[1, 14] did not provide an actual value for EER at 40sec. These values were interpreted from the Tippett plots and tables provided in the papers.*

The voiced duration of the current study was altered to 40 sec to match the study [14] which found the mean available voiced duration of 41 secs. As demonstrated by Table 2, the male Australian speakers' long term F0 are not significantly superior or inferior to other languages as an individuating FVC feature. However, it needs to be emphasised that we cannot directly compare these experiments as they were not carried out in an identical way.

6. Conclusion and Future Studies

This research has dealt with the discriminating potential of long term F0 in male Australian English speakers in the context of the LR-based FVC. The strength of evidence (LRs) was estimated using the multivariate kernel density (MVDK) formula together with the F0 based features extracted by the six parameter method.

It has been demonstrated that long term F0 is useful for FVC cases in (Australian) English, in particular when the sample size is relatively large. However, the samples were collected under laboratory conditions, and it would be useful to be able to test the efficacy of long term F0 under more realistic conditions.

It has been shown that different register types affect the performance of the FVC system in that the FVC system performs better with data elicited in a casual setting (e.g. CNV) than a formal setting (INT). The degree of expressiveness associated with different registers has been discussed as a possible reason for this, but a further study is required for confirmation.

Although a trade-off between the system validity and reliability has been reported in some previous studies, the results of the current study indicate that an FVC system can be improved both in its validity and reliability with an appropriate amount of data. Yet, again this also warrants further studies.

7. References

[1] Y. Kinoshita, S. Ishihara, and P. Rose, "Exploring the discriminatory potential of F0 distribution parameters in traditional forensic speaker recognition," *International Journal of Speech Language and the Law,* vol. 16, pp. 91-111, 2009.

[2] J. Li and P. Rose, "Likelihood ratio-based forensic voice comparison with F-pattern and tonal F0 from the Cantonese /ɔy/ diphthong," in *14th Australasian International Conference on Speech Science and Technology*, 2012, pp. 201-204.

[3] G. M. Morrison, 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,* vol. 44, pp. 155-167, 2012.

[4] Australian Bureau of Statistics. (2016). *Safety and Justice*. Available: http://www.abs.gov.au/ausstats/abs@.nsf/Lookup/by%20Subject/4125.0~Feb%202016~Main%20Features~Safety%20and%20Justice~4411

[5] F. Nolan, *The Phonetic Bases of Speaker Recognition*. Cambridge: Cambridge University Press, 1983.

[6] E. Enzinger and G. S. Morrison, "Empirical test of the performance of an acoustic-phonetic approach to forensic voice comparison under conditions similar to those of a real case," *Forensic Science International,* vol. 277, pp. 30-40, Aug 2017.

[7] C. E. Williams and K. N. Stevens, "Emotions and Speech - Some Acoustical Correlates," *Journal of the Acoustical Society of America,* vol. 52, pp. 1238-&, 1972.

[8] B. Robertson and G. A. Vignaux, *Interpreting Evidence: Evaluating Forensic Science in the Courtroom*. Chichester: John Wiley, 1995.

[9] C. G. G. Aitken and D. Lucy, "Evaluation of trace evidence in the form of multivariate data," *Journal of the Royal Statistical Society, Series C (Applied Statistics),* vol. 53, pp. 109-122, 2004.

[10] G. S. Morrison, "Measuring the validity and reliability of forensic likelihood-ratio systems," *Science & Justice,* vol. 51, pp. 91-98, Sep 2011.

[11] B. Nair, E. Alzqhoul, and B. J. Guillemin, "Determination of likelihood ratios for forensic voice comparison using principal component analysis," *International Journal of Speech Language and the Law,* vol. 21, pp. 83-112, 2014.

[12] S. Ishihara, "Sensitivity of likelihood-ratio based forensic voice comparison under mismatched conditions of within-speaker sample sizes across databases," *Australian Journal of Forensic Sciences,* pp. 1-16, 2017.

[13] D. Gibbon, "A new look at intonation syntax and semantics," in *New Linguistic Impulses in Foreign Language Teaching*, R. A. James and P. Westney, Eds., ed: Tuingen: Narr, 1981, pp. 71-98.

[14] R. Zheng and P. Rose, "Likelihood ratio-based forensic voice comparison with Cantonese short-term fundamental frequency distributino parameters," in *14th Australasian International Conference on Speech Science and Technology*, 2012, pp. 153-156.

Cross-language categorisation of monosyllabic Thai tones by Mandarin and Vietnamese speakers: L1 phonological and phonetic influences

Juqiang Chen[1], Catherine T. Best[1,2], Mark Antoniou[1], Benjawan Kasisopa[1]

[1]The MARCS Institute, Western Sydney University, Australia
[2]Haskins Laboratories, New Haven CT, USA
J.Chen2/C.Best/M.Antoniou/B.Kasisopa@westernsydney.edu.au

Abstract

This study explores the influences of listeners' native tone inventory on cross-language tone perception. Mandarin, Northern Vietnamese and Southern Vietnamese listeners (n = 13 per group; naive to Thai) categorised Thai tones into their native tone categories. Results show that all three groups categorised most Thai tones into their native tone categories. Their performance suggests that they attended to the phonetic details of the non-native tones, and that their assimilation patterns were influenced by the organisation of their native tone phonological systems as well as by the phonetic distances between native and non-native tones.

Index Terms: cross-language perceptual assimilation, non-native lexical tones

1. Introduction

Lexical tones are pitch variations that change the meanings of words. Although more than 60% of the languages in the world are tonal [1], non-native tone perception [2]–[5] is understudied relative to perceptual studies on non-native consonants [6] and vowels [7]. Moreover, most studies have used Mandarin or Cantonese as target languages and tested listeners of non-tone languages, such as English or French [5]. Southeast Asian tone languages (e.g., Thai, Vietnamese) are seldom investigated. How native speakers of these languages perceive each other's tones is largely unknown.

To examine this, we drew upon the Perceptual Assimilation Model (PAM), which assumes that perception of non-native phonemes depends on their similarities to the native phonemes that are in closest proximity to them in native phonological space [8]. We used cross-language tone categorisation tests to determine how the listener's native tone language affects assimilation of non-native tones to the native system.

1.1 Non-native tone assimilation

According to PAM [8], both phonetic similarities and L1 phonology affect how non-native tones are perceived. To predict likely phonetic effects for cross-language tone assimilations, we used Chao's notation, in which F0 height at tone onset and offset is referenced by numbers 1-5 ranging from low to high [9]. For example, in [10] Cantonese speakers categorised the Mandarin (M) high-level tone M55 into the phonetically similar Cantonese (C) category high-level C55. However, the Mandarin high-falling tone M51 was also categorised into Cantonese C55, i.e., not a falling tone as would be predicted simply from its pitch contour. The authors' interpretation was that the Cantonese listeners perceived the Mandarin high-falling tone as Cantonese C55 because they perceived it as similar to the Cantonese high-falling C53 allotone of C55. In other words, in this case they relied on their native tone phonology. Thus, as PAM predicts, for non-native tone assimilation listeners tap into both L1 phonetic details and L1 phonological tone space with respect to accessing information about allophonic variation of tones.

Similarly, Mandarin listeners categorised the Thai (T) high-level tone T45 to the phonetically similar Mandarin mid-rising tone M35 [2], [11]. However, they assimilated the Thai mid-level tone T33 to the phonologically similar Mandarin high-level tone M55 and the Thai low-level tone T21 was categorised by participants in [2] as the Mandarin high-level tone M55, which is the only level tone in Mandarin but has a different pitch height. In these cases, they used the same abstract phonological type across the non-native language target tone and their L1 tone (both phonologically 'level' tones) even though the phonetic pitch height was dramatically different. Mandarin listeners in [11], however, categorised T21 as the Mandarin high-falling tone (51). They chose the same phonetic contour (falling pitch) even though the abstract phonological type and the starting pitch height differed. Thus participants in the two studies seem to have based some choices on the phonetic level but other choices on the phonological level of their L1 tone systems. This necessitates further investigation of the phonetic versus phonological basis of non-native tone assimilation.

The present study tested how Mandarin (4 tones), Northern Vietnamese (NV, 6 tones) and Southern Vietnamese (SV, 5 tones, with a tone merger) listeners categorised the tones of the unfamiliar target language, Thai (5 tones), into their native tone inventories. One motivation to use Mandarin, NV, and SV listener groups is that the differences in the number of tones in the listeners' native language phonological tone inventories could affect non-native tone assimilation. This is important to address given that prior findings on whether tone inventory size benefits non-native tone perception [12] or not [13] are inconsistent.

1.2 Characteristics of Mandarin, Thai, and Vietnamese tones

In the present study, in order to develop PAM-motivated phonetic and phonological predictions for language-specific tone assimilations, we will indicate the phonological features of tone types in terms of height (high, mid, low) and pattern (level or rising, falling, falling-rising or rising-falling contour), while using Chao values [9] to give a priori phonetic descriptions of the tones in each language. Thai, the target language, has three level tones (high-level T45, mid-level T33, low-level T21) and two contour tones (characterised as mid-falling-rising T315 and low-rising-falling T241 in [2] but respectively as low-falling-rising 214 and high-falling 41 in

[11]). In the present study, isolated citation form syllables were used; thus our hypotheses were based on [2].

It is generally agreed that Mandarin has four tones in citation form: a high-level tone M55; a mid-rising tone M35; a low-falling-rising tone M214; and a high-falling tone M51.

NV has six tones. Nguyen and Edmondson [14] measured them on the syllable /ta/ produced by six informants and assigned Chao numbers to each. NV has one mid-level tone NV33 called *ngang*. *Sắc* is mid-rising NV35, while *ngã* is mid-falling-rising NV315. Low-falling NV21,called *huyền,* starts low and moves slowly downward, while mid-falling NV32 *nặng* is shorter in duration. Mid-falling-rising NV313 *hỏi* starts somewhat higher than *huyền* and drops rather abruptly, followed by a moderate rise at the end in citation form.

SV has five tones. SV *ngang, sắc* and *huyền* (referred to as SV33, SV35 and SV21, respectively) are phonetically very similar to their NV counterparts. SV *nặng* is low falling-rising SV212, while SV merges *hỏi* and *ngã* tones into a single low-falling-rising tone SV214 [14, 15].

1.3 Predictions

Extending PAM [8] principles to non-native tone assimilation by listeners of other tone languages, we posit that a non-native tone can be assimilated to a native tone category as an ideal or acceptable or deviant exemplar, i.e., Categorised. Otherwise, a non-native tone can be assimilated within native phonological space but not as a clear example of any particular native category, thus Uncategorised. Our predictions concerning how tone language experience affects perceptual assimilation are based on phonetic similarities (i.e., Chao notations) and/or phonological similarities (e.g., tone types: height and pattern) between the target and the native phonemes that are in closest proximity to them in native phonological space.

Thai has three level tones whereas Mandarin has only one. Thus, Mandarin listeners may categorise the three Thai level tones into the single native Mandarin level tone M55 because they share the phonological type "level tone." However, as none of the Thai level tones matches M55 phonetically, this could lower the percent categorisation of T33 to M55, and shift categorisation of T45 and T21 to the "rising contour" type of M35, but also lower the percentages of those categorisations. Thai mid-falling-rising T315 should be categorised phonologically as the "falling-rising" M214. However, because M214 is realized as 21 when followed by another tone in a two-syllable word, i.e., it is an allotone of M214, this may lead to phonological categorisation of T21 as M214. Low-rising-falling T241 should be Uncategorised, though, as Mandarin lacks the rising-falling phonological type.

As NV and SV each have only one mid-level tone (NV33, SV33), it is likely that T33 will be strongly categorised to their native level tone both phonologically and phonetically. The phonetic mismatch of the other Thai level tones, T45 and T21, from NV/SV33 may either result in weaker categorisation to NV/SV33 level tone, or to lower categorisation of T45 to native mid-rising tones NV35/SV35 and strong categorisation of T21 to native low-falling tones NV21/SV21. Similarly, T315 should be strongly categorised to NV315 and less strongly categorised to SV214. However, T241 should be Uncategorised because NV and SV, like Mandarin, both lack any phonologically rising-falling tones.

2. Method

2.1 Participants

Participants were 13 native speakers each, of Mandarin (M_{age} = 30.5 years), Northern Vietnamese (M_{age} = 20.3 years), and Southern Vietnamese (M_{age} = 23.2 years). According to a background questionnaire given before the test, none had more than two years of formal musical training, which is important because musical experience could influence tone perception. All had normal hearing and none had experience with Thai.

2.2 Stimulus materials

Two syllables (/ma/, /mi/) were chosen because they are real words for each native tone in Thai, Mandarin and Vietnamese. Thus naïve listeners were able to categorise the Thai tone stimuli into their native phonological systems, as lexical items. The target Thai syllables were each read several times by two female native Thai speakers. These informants had no experience with other tone languages. Two tokens of each target item that were judged to be correct and most natural-sounding to a third native Thai speaker were selected for use as perceptual stimuli.

2.3 Procedures

Participants were tested individually in a quiet room (e.g., testing booth at Western Sydney University, library study booths at Nanjing University, Macquarie University, University of Technology Sydney). Stimuli were presented on a Dell Latitude 7280 laptop running E-Prime Professional 2. Auditory stimuli were presented via Sennheiser HD 280 Pro headphones at 72 dB SPL.

Before the test session, participants completed 20 practice trials. The categorisation task had 120 trials (2 speakers × 5 tones × 2 tokens × 2 syllables × 3 repetitions). On each trial, the stimulus token was presented and listeners made a forced-choice categorisation judgment to their native tones via a key press. Mandarin participants chose from four Pinyin options, and Vietnamese speakers from six Vietnamese transcriptions. Participants were asked to respond as quickly as possible within a 3s response period. Choices of *hỏi* and *ngã* tones (orthographically different) by the SV group were combined for data analysis as they are phonologically merged rather than contrastive in SV, as reflected in non-systematic choices between these options, unlike the NV group's distinct choices.

2.4 Categorisation criteria

Although in some studies [2], [16], [17], "Categorised" means one native category has to be selected 50% or 70% of the time for a given nonnative feature, we followed the tone categorisation criteria used in [5]: First, a given native tone must be selected significantly more than chance level, which was 25% (1/4) for Mandarin, 16.7% (1/6) for NV, and 20% (1/5) for SV speakers. Second, one native tone category must be chosen significantly more often than choices of any other response categories. This method is sensitive to variation among different native tone systems as it takes into account the number of tone categories in a particular language.

3. Results

3.1 Tonal categorisation of Mandarin, NV and SV listeners

First, we determined whether categorisations were above

chance level via a series of t-tests. Table 1 shows choices for each Thai target, with those that are significantly above chance for Mandarin (M, 25%), Northern Vietnamese (NV, 16.7%) and Southern Vietnamese (SV, 20%) in bold. Assimilations that met the Categorisation criteria are denoted in Table 1 with an asterisk.

Having established which categorisations were above chance, we next examined differences in categorisation/ assimilation patterns across the groups. Since the number of native tone choices differed for each language group, this did not permit a repeated measures ANOVA to compare between language groups. However, assimilation patterns are informative enough for comparison. Chi-square tests within each listener language revealed that native language choices were not evenly distributed across the Thai targets, suggesting associations between specific Thai targets and native choices (Mandarin, $\chi^2(12) = 2170.1$, $p < .001$; NV, $\chi^2(20) = 2136.9$, $p < .001$; SV, $\chi^2(16) = 2002.7$, $p < .001$).

For each group, one-way ANOVAs for the five Thai target tones were carried out to determine whether listeners assigned different native tone category labels to the Thai targets.

Table 1. *Categorisation of Thai tones into Mandarin (M), Northern Vietnamese (NV) and Southern Vietnamese (SV) tone categories*[a].

Choices	Thai targets				
	T45	T33	T21	T315	T241
M55		**92.6***	10.9		**48.0**
M35	**88.3***			**79.7***	
M214	10.1		**62.5***	20.3	
M51			25.7		**51.0**
Assim:	***C***	***C***	***C***	***C***	***U***
NV33		**29.8**			**92.9***
NV35	**53.1***			**47.7***	
NV315	21.6			**26.5**	
NV32					
NV21		**67.3***	**76.9***		
NV313	14.8		17.6	22.9	
Assim:	***C***	***C***	***C***	***C***	***C***
SV33		**33.8**			**83.5***
SV35	29.8			10.7	
SV214	31.2			**80.1***	
SV212	30.8				
SV21		**56.8***	**77.3***		
Assim:	***U***	***C***	***C***	***C***	***C***

[a]Table notes: Categories in bold are choices that were significantly above chance (25% for Mandarin, 16.7% for NV, 20% for SV). Assimilations: C = Categorised, U = Uncategorised. "*" = Categorised tone. Only choices > 10% are presented.

For Mandarin listeners, significant differences among Mandarin choices were found for all Thai target tones: T45, $F(3) = 234.2$, $p < .001$; T33, $F(3) = 302.0$, $p < .001$; T21, $F(3) = 11.3$, $p < .001$; T315, $F(3) = 116.5$, $p < .001$; T241, $F(3) = 19.8$, $p < .001$. Post-hoc Tukey tests revealed that for T45, the mean percentage of M35 choices was significantly greater than those of M55, M214 and M51, $p < .001$. For T33, the mean percentage of M55 choices was significantly greater than other Mandarin tones, $p < .001$. For T21, choice of M214 was significantly greater than other Mandarin tones, $p < 0.01$. For T315, choice of M35 was significantly greater than other Mandarin tones, $p < .001$. For T241, the choice of M55 and M51 were not significantly different from each other. Thus, T45, T33, T21 and T315 were all Categorised by Mandarin listeners, while T241 was Uncategorised.

For NV listeners, significant differences in NV choices were found for all Thai target tones: T45, $F(5) = 16.1$, $p < .001$; T33, $F(5) = 106.8$, $p < .001$; T21, $F(5) = 106.6$, $p < .001$; T315, $F(5) = 19.9$, $p < .001$; T241, $F(5) = 457.8$, $p < .001$. Post-hoc Tukey tests revealed that for T45, the mean percentage of categorisations as NV35 was significantly greater than other NV tones, $p < .001$. For T33, choice of NV21 was significantly greater than that of other NV tones, $p < .001$. For T21, choice of NV21 was significantly greater than all other choices, $p < .001$. For T315, choice of NV35 was significantly greater than the other choices, $p < .01$. For T241, choice of NV33 was significantly greater than all the other NV tone choices, $p < .001$. Thus, all Thai tones met the Categorised criteria for NV listeners.

Furthermore, for SV listeners, significant differences in SV choices were also found for all Thai target tones: T45, $F(4) = 4.2$, $p < .01$; T33, $F(4) = 39.6$, $p < .001$; T21, $F(4) = 34.1$, $p < .001$; T315, $F(4) = 34.8$, $p < .001$; T241, $F(4) = 76.8$, $p < .001$. Post-hoc Tukey tests revealed that for T45, no single native choice was significantly higher than others, thus T45 was Uncategorised. All other Thai tones were Categorised by the SV group. For T33, the mean percentage of SV21 choices was significantly greater than others, $p < .001$. For T21, SV21 was significantly greater than all the other choices, $p < .001$. For T315, choice of SV214 was significantly greater than other tones, $p < .05$. For T241, choice of SV33 was significantly greater than all other SV choices, $p < .001$.

4. Discussion

Generally, our three groups of native tone language listeners (Mandarin, NV, SV) categorised the non-native Thai tones into native categories, except for two cases. T241 was not categorised into a single Mandarin tone category, as predicted by PAM due to the lack of a rising-falling tone in Mandarin phonology. Moreover, the choices were split across two Mandarin categories, M55 and M51, suggesting that listeners were inconsistent in attending to pitch height (phonetic details) versus pitch contour (phonological type).

T45 was not categorised into a single SV category. Three tones (SV35, SV214, SV212) were chosen but none were chosen significantly above chance nor significantly more often than others. The reason could be that there is no high-level tone in SV for phonological assimilation and there is no good phonetic match (i.e. SV35 has a lower onset). Interestingly, NV listeners chose NV35 significantly more than other NV tones for T45. This suggests that NV listeners were not constrained by the phonological category (high-level vs. mid-rising) and ignored the phonetic onset pitch difference. NV315 and NV313 were occasionally chosen, which suggests that Vietnamese listeners may match the rising contour of T45 to the middle-to-final rising contour of NV313 and NV315. Given that SV merges tones 313 and 315, SV listeners may be insensitive to the subtle difference in middle-to-final rising contours, and therefore cannot decide which category to put T45 into (SV35, final 14 in SV214, or final 12 in SV212).

For Thai level tones, Mandarin listeners categorised T33 into the phonologically similar Mandarin level tone M55 as PAM predicted. However, they categorised T45 into M35 as rising pitch despite the phonetically differing onset pitch. Interestingly, and as we predicted, T21 was categorised as M214, which is often realized as 21 in two-syllable Mandarin words. This suggests that Mandarin listeners' assimilations were phonologically influenced by the allophonic organization of their native tone system.

For both NV and SV listeners, low-level T21 was phonologically and phonetically categorised as low-falling NV21 and SV21 respectively instead of being categorised as level tone NV/SV33. However, T33 was also not categorised as the phonologically and phonetically similar mid-level NV33 nor SV33, but instead also as low-falling NV21 and SV21. This may reflect a phonetic inconsistency in use of Chao tone numbers, i.e., T33 may differ phonetically from NV33 and SV33, as it appears that Chao number 3 in Thai corresponds more closely to Chao number 2 in NV and SV. This needs further investigation.

Thai mid-falling-rising tone T315 was categorised by Mandarin listeners as more like mid-rising M35 than phonological low-falling-rising M214. Mandarin listeners tended to simplify the complex T315 by attending only to the middle-to-final pitch contour (15). Given that M214 may occur often as 21 in sentences, it may not be an ideal choice for T315. For NV listeners the same is true: the simple contour NV35 was chosen more than other candidates (NV315, NV313). SV listeners chose SV214 more than a simple rising tone SV35. Given that SV214 merges the 315 and 313 tones of NV, this merged tone category is more prominent in the SV phonological space. If we assume Thai tone pitch height in Chao numbers may be perceived as lower by Vietnamese speakers, SV listeners may not distinguish the difference in pitch height between T315 vs. SV214.

Thai low-rising-falling tone T241 was uncategorised by Mandarin listeners. But both NV and SV listeners perceived it as a native mid-level tone, NV33 and SV33. This is in line with the tendency to categorise complex non-native tones to simpler native contour tones. In addition, NV and SV rising tones (35) are higher than the 24 portion of T241, and falling tones lower (32/21) at their onsets than the 41 portion of T241, thus offering only a mediocre phonetic match to either portion of T241.

In summary, first, phonological constraints are strong (e.g. T33 → M55, phonological type "level tone" supersedes phonetic pitch differences) but can be affected by cross-language phonetic dissimilarities (high level T45 → mid-rising M35, N35, instead of level tones). Second, listeners have access to allophones of a native tone if it is phonetically similar to the non-native tone. Third, complex non-native contours are likely to be perceptually simplified and hypothetically the middle-to-final part is attended.

In the present study, we used Chao notation for comparing phonetic details of lexical tones between languages. However, Chao notation does not capture acoustic details such as duration, rate of pitch change and phonation types, which may affect how listeners assimilate non-native tones. Moreover, Chao numbers are normalized within a given language, not between languages. Thus, a Chao number of 3 in one language may reflect a different actual F0 value in another language, as we have mentioned above for 3 in Thai relative to 2 in Vietnamese. The magnitude of the F0 difference may be large enough be perceptually distinguishable. Therefore, more detailed acoustic information should be taken into consideration in future research on phonetic-level effects in cross-language tone perception.

5. Conclusion

Our findings supported most PAM-motivated predictions. The listeners did not simply match target tones with corresponding native tones based only on similar phonetic details. Instead, phonological constraints are strong but can be affected by cross-language phonetic dissimilarities. Positional allophonic information affects listeners' categorisations of non-native tones into native tone categories. Naïve listeners may simplify the complex contour tones by attending to the middle-to-final part of the contour. Thus, it is essential to consider language-specific tone assimilation patterns as these will determine how naïve listeners will discriminate particular non-native tone contrasts. These findings have ramifications for theories of tone perception as well as pedagogical implications for second language lexical tone training [18].

6. References

[1] Yip, M., Tone, Cambridge University Press, 2002.

[2] Reid, A. et al., "Perceptual assimilation of lexical tone: The roles of language experience and visual information," Atten. Percept. Psychophys., 77:571–591, 2015.

[3] Burnham, D. et al., "Universality and language-specific experience in the perception of lexical tone and pitch," Appl. Psycholinguist., 36:1459–1491, 2015.

[4] So, C. K. and Best, C. T., "Cross-language perception of non-native tonal contrasts: effects of native phonological and phonetic influences," Lang. Speech, 53:273–293, 2010.

[5] So, C. K. and Best, C. T., "Phonetic influences on English and French listeners' assimilation of mandarin tones to native prosodic categories," Stud. Second Lang. Acquis., 36:195–221, 2014.

[6] Best, C. T., McRoberts, G. W. and Goodell, E., "Discrimination of non-native consonant contrasts varying in perceptual assimilation to the listener's native phonological system," J. Acoust. Soc. Am., 109:775–794, 2001.

[7] Tyler, M. D., Best, C. T., Faber, A. and Levitt, A. G., "Perceptual assimilation and discrimination of non-native vowel contrasts," Phonetica, 71:4–21, 2014.

[8] Best, C. T., "A direct realist view of cross-language speech perception.," in W. Strange, [Ed.], Speech perception and linguistic experience: Issues in cross-language research, 171–204, York Press, 1995.

[9] Chao. Y.R., "A system of tone-letters," Maitre Phon., 45:24–27, 1930.

[10] So, C. K., "Cross-language categorization of monosyllabic foreign tones : Effects of phonological and phonetic properties of native language," in T. Stolz, N. Nau, and Stroh, [Eds.], Monosyllables : From Phonology to Typology, 55–69, 2012.

[11] Wu, X., Munro, M. J. and Wang, Y., "Tone assimilation by Mandarin and Thai listeners with and without L2 experience," J. Phon., 46:86–100, 2014.

[12] Qin, Z. and Mok, P. P.-K., "Discrimination of Cantonese tones by Mandarin, English and French speakers.," presented at the the workshop on the psycholinguistic representaion of tone, Hongkong, 2011.

[13] Chiao, W.-S., Kabak, B. and Braun, B., "When more is less: Non-native perception of level tone contrasts," in Proceedings of the Psycholinguistic Representation of Tone Conference, 2011.

[14] Nguyen, V. L. and Edmondson, J. A., "Tones and voice quality in modern northern Vietnamese: instrumental case studies," Mon-Khmer Stud., 28:1–18, 1998.

[15] Brunelle, M., "Tone perception in Northern and Southern Vietnamese," J. Phon., 37:79–96, 2009.

[16] Antoniou, M., Tyler, M. D. and Best, C. T., "Two ways to listen: Do L2-dominant bilinguals perceive stop voicing according to language mode?," J. Phon., 40:582–594, 2012.

[17] Bundgaard-Nielsen, R. L., Best, C. T. and Tyler, M. D., "Vocabulary size matters: The assimilation of second-language Australian English vowels to first-language Japanese vowel categories," Appl. Psycholinguist., 32:51–67, 2011.

[18] Best, C. T. "The diversity of tone languages and tonality in non-tone languages: How well does perceptual research reflect this diversity?" Front. Psychol., in press 2018.

Pitch accent movements in expressive speech

Grażyna Demenko

Institute of Linguistics, Adam Mickiewicz University, Poznań, Poland

lin@amu.edu.pl

Abstract

This contribution presents the description of pitch accent movements and their interpretation for the needs of processing expressive speech. Based on chosen perceptually utterances from 40 speakers recordings from the Police 997 emergency call center, a few hundred pitch accent shapes based on F_0 changes on accented and postaccented syllables have been analyzed. Vocal register width and position on the scale of fundamental frequency correlated with emotional state of speaker. Statistics have shown that pitch accent movements are dependent on emotions in expressive speech. In the case of extreme stress linguistic functions of intonation have been significantly reduced.

Index Terms: pitch accent movement, emotions, expressive speech

1. Introduction

Use of prosodic information in speech technology calls for clearly defined basic requirements of contemporary prosodic transcription: (a) the prosody description should be well-grounded theoretically, (b) prosody forms should consider representation of pitch accent with reference to: position and type of F_0 change, size and change rate, synchronization of pitch movement with syllables and phrase structure, (c) prosodic functions should be possibly formulated on all language levels.

Currently, the most significant problems in modeling prosodic structures concern formulation of acoustic determinants of accent and contextual description of accent that would allow for analysis of various functions of intonation, e.g. [1]. From the point of view of intonation functions analysis, modeling the register width and position on the scale of fundamental frequency in relation to the whole range of the speaker's pitch height is crucial.

For description of variation in pitch range the terms "key" (for discourse purposes) and "register" considered as an overall (upward and downward) shifting of the whole pitch range within which the speaker is speaking are the most frequently used [2,3]. In the literature on the subject the vocal register is one of the least clear concepts [4]. The most often quoted definition originates from Hollien [5]: „(...) register is defined from acoustical or physiological point of view". The function of register is to signal the emotional state of a speaker, e.g. tension, stress or anger. High register may signal a threat or social politeness, e.g. [6].

Discussing the importance of an elevated vocal effort and other factors, which can influence the whole range of F_0 parameter changes, a variability model has been suggested, in which two dimensions are distinguished: level and range.

A hypothesis has been put forward that, apart from differentiating distinctive levels, such as L, M, H, there may be a parameter describing the range that indicates a shift on a scale of all the pitch movement in the current study the term "register" as defined by Hollien will be used. A commonly used technique for pitch range changes discrimination is normalization related to the distance between the lowest and the highest F_0 value reached by a speaker. A different possibility for fundamental frequency height change normalization is its relative location (in relation to the preceding pitch), e.g. [8]. Analyzing various methods of pitch changes normalization [9], it should be noted that an objective scaling of fundamental frequency can be ensured by a physical scale (of a limited range for speech), so that it allows for unambiguous indication of the individual range of changes for pitch height and location on the scale, e.g. [10,11]. Absolute F_0 is the most important information for indication for both pitch level and speaker sex.

The problem of pitch accent movement has also been addressed by many studies e.g. [12]. It has been concluded that accent realization may indicate not only speaker's linguistic competence but also speaker's emotional state such as stress or anger or fear.

Different emphases have been noted in early studies such as [12], where three forms have been pointed out: a) upward obstruction of F_0 (rise/fall or jump/drop), b) steep fall plus drop before the next syllable (jump/fall drop), and c) the secondarily emphasized syllable interrupts the more rapid fall of the surrounding of non-emphasized syllables.

The discussion of the functions of intonation for needs of human computer voice interfaces is the subject of the following section. The formal description of the accent should take into account all the most important functions of intonation. In particular, the least studied neuropsychological functions, which require much more sophisticated accent description and have been not used so far in the intonation modeling. A particular attention to this issue is included in this study. The third section is concerned with formalization of pitch accent representation, and the fourth section deals with the methodology of pitch movement analysis in expressive speech, especially speech under stress.

2. Functions of intonation

There is no consensus as to the priorities of multiple functions of intonation. Some researchers emphasize the grammatical functions, whereas others underline that grammatical functions are secondary to emotional functions [13,14,15].

Linguistics functions. The most important factors for linguistic motivated features are: speaker's attitude, discourse condition and thematic accent placement.

Discourse functions. The discourse function regulates conversational behavior, it provides information about what the speaker is doing in speaking, that is whether he/she is questioning, advising, encouraging, disapproving.

An appropriate pitch movement (fall/rise/level) signals if the

phrase is closed, finished, definitive/open, neutral.
The F_0 register signals differences in emphasis of finality or nonfinality.

Sociolinguistic functions. The functions of intonation on the sociolinguistic level provide information about regional varieties, sociocultural background and socioeconomic status, and it is especially useful for speaker characterization and speech recognition when taking into account pronunciation variants, on both segmental and suprasegmental levels.

Psycholinguistic and socio-psychological functions. Certain emotional states which can be controlled by the speaker to some extent, are often correlated with physiological states which in turn have quite mechanical, and thus predictable effects on speech, especially on its prosodic structure. For instance, when a person is in a state of anger, fear or joy, the sympathetic nervous system is aroused and the speech becomes loud, fast and enunciated with a strong high-frequency energy [16]. When one is bored or sad, the parasympathetic nervous system is aroused, which results in a slow, low-pitched speech with little high-frequency energy. Furthermore, the fact that these physiological effects are rather universal means that there are common tendencies concerning the acoustic determinants of basic emotions [17]. The patterns of complex pitch movement: rise-fall/fall-rise mostly correlate with emotions and emphasis (e.g. rise-fall: possible correlation – approval, admiration; fall-rise: possible correlation - astonishment, disbelief).

Psychological functions. It has long been known that, quite apart from what is said, a speaker's voice conveys considerable information about the speaker, and that listeners utilize this information in evaluations of speaker's attributes, the speaker's personality traits, speaker charisma, dominance, subjectivity, opinion and sentiment.

***Neuropsychological (neurocognitive) functions*.** Stress in relation to speech needs to be carefully defined, to distinguish it from the meaning of stress associated with an accented syllable in linguistics. Therefore, the speech under stress, which is less ambiguous, will be used. Stress produced in response to the occurrences in the people's surroundings, perceived by them as unusual and impossible to control, belongs to third-order stressors, psychological ones [18]. These kinds of stressors have their effect at the highest level of speech production and cause extreme changes both in segmental and suprasegmental speech structures [19].

3. Accent shape annotation

An accented syllable initiates intonation changes on the subsequent syllable/syllables and determines the tone type (static/dynamic) and tone configuration (falling/rising/flat), e.g. [20,21,22]. Dynamic tone carriers are those vowels and sonorants on which the F_0 change is greater than 2.5 semitones (ST), rise for L, fall for H. Static tone carriers are those vowels and sonorants on which there is no F_0 change or the change is less than 2.5 ST (marked as L_ if the tone on the next vowel is higher or H_ if the tone on the next vowel is lower).

A structure, where the course of F_0 parameter on accented syllable is lower than the tone of the next syllable is assumed to be a rising course, and a structure where F_0 on the accented syllable is higher than the tone of the next syllable is a falling course. First letter represents the course of F_0 on an accented vowel, the next letter stands for the course of F_0 on the vowel directly following the accented vowel. If the basic tone on an accented and postaccented vowel is close to even (within the range of +/-2 ST), then such a tone is marked as even F. If the F_0 course within the accented syllable can be approximated by a quadratic function than such a tone change is marked with L&H. If the tone on the accented vowel initiates F_0 rise/fall on a few subsequent syllables than such a tone is marked as ascending L*/descending H*. For annotation of accented structures in expressive speech the description formulated in [24] has been used:

[P][Z][N] [SA{!} {PA}][S]

where:

[P] within phrase position marker

[Z] range. Referring to the whole structure (comprising accented and not accented syllable): A – range less than or equal to 3 ST, B – range within: 4 ST <= 6 ST, C – range within: 7 ST <= 9 ST, D – range within: 10 ST <= 12 ST, E – range greater than 12 ST. The dynamics of 3 types of accents have been distinguished: extradynamic accent : L!, H! – changes greater than 4.5 ST, dynamic accent: L, H – changes within 2 – 4.5 ST, static accent: L_H_ – changes smaller than 2 ST.
Symbol ! refers to the dynamics of changes on the accented syllable. The range refers to calculating the difference between F_{min} and F_{max} of the fragment of the utterance that contains the analyzed accent structure.

[N] accent position on a frequency scale is represented by a number N expressed in ST (N = 3, 6, 9, 12, 15, 18, 21, 24, 27, 30, 33, 36). Semitones are related to F_{min}= 60 Hz. In the view of very intricate (and counterintuitive) rules needed to interpret tones that are defined relatively, the accent structure position on the frequency scale has been determined in absolute values.

[S] steepness (DF/DT{s}). Expressed in semitones per seconds over the accented vowel (and neighboring sonorants).

[SA{!} {PA}] accent structures (LH, LL LH_ , HL,HH_, HL_ , L_H, L_L, L_H_ H_L_H_H, F, L&H L*, H*), where the first letter indicates the type of F0 course on the accented syllable (A) and the second one on the postaccented syllable (PA).

4. Methodology

4.1. Speech Data

Out of 20 000 recordings selected automatically according to their duration (dialogs shorter than 3-4 seconds were omitted) from the database, the recordings of 40 out of a few hundred speakers were chosen for analysis for acoustic evaluation, the basis for selection being a perceptual assessment. It was assumed that a particular situation which makes a person call 997 determines more or less notably the way he/she speaks.

Four situational contexts were taken into account [24]:

1). Direct life hazard – extreme stress (**ES**). A situation in which the person calling or his/her friends/relatives, etc. are in a **direct threat of losing their life** produces extreme stress.

2). Indirect life/health hazard, a threat of losing property – **limited stress** (**S**). A situation in which the person calling or his/her friends/relatives are in an indirect threat of losing their life, health or property produces stress whose intensity depends on the individual characteristics of a person. Different emotions can be produced by a speaker like fear, anger, disgust.

3). Tragedies of life – **depression (D).** A situation in which the persons calling or his/her friends/relatives, etc. are potentially threatened by losing their life/health as a result of a terminal illness (or planned suicide, etc.) also other misfortunes in life.

4). Utterance which does not carry emotions – **neutral stress** (**N**). This utterance type is characteristic of informative calls, when the person calling wants to get some information concerning an address, a telephone number, etc. The material was divided into four groups, each group contained 10

speakers: **ES, S, N, D**. For pitch movement analysis the PitchLine software presented in [23] was used.

4.2. Pitch movement analysis

The following cases have been analyzed: 1) dynamic changes in register position during the utterance, 2) relative constant register position within the utterance, 3) relative constant register position within the utterance, width register variable and 4) changes in pitch register width and its position.

1). Dynamic changes in register position

Fig.1 illustrates an utterance of a female marked by an extreme stress (**ES**). At the end of the utterance the pitch is high with small variability. In the case of extreme stress, pitch accents occurred mainly in the initial fragments of the speech (in the lower range of F_0 changes) and in 90% they were falling or level accents.

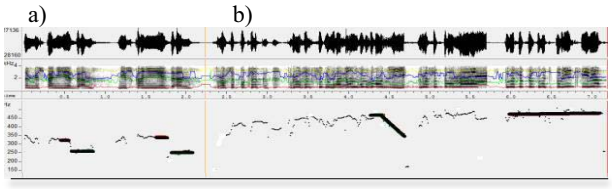


Figure 1: *A gradual stress increase in the utterances: a) Someone is entering the apartment (F_{min} =220Hz), b) he is somewhere [here] - direct threat (F_{min} =345 Hz). Three accents (two H_L) and one (H_L) were recognized.*

2). Relative constant register position within the utterance

Fundamental frequency contours in speech produced by a person in depression (**D**) are relatively flat with small variability (2-3 ST). It was not possible for speech samples representing deep depression to detect pitch accents (Fig. 2).

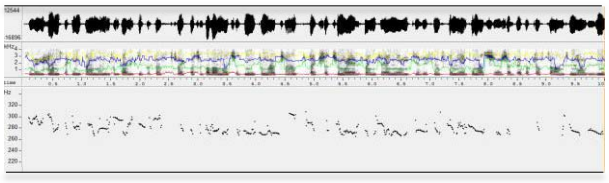


Figure 2: *F_0 contour for an utterance produced by a person suffering from depression: I left for Belgium with one man and now he's detaining me, I'm being held. (F_{max} =302 Hz, F_{min} =240 Hz).*

Fig.3 shows also F_0 contour relative flat (2-3 ST), in speech produced by a person in depression with changing register (after an accented syllable the next syllable starts in vocal fry register see e.g.[25]).

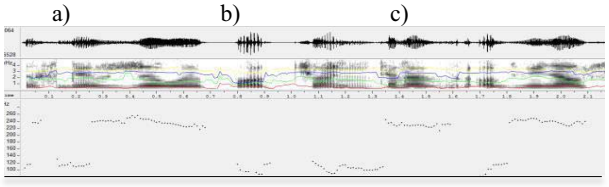


Figure 3: *F_0 contour for an utterance produced by a person suffering from depression: a) my brother wants, b) to cut his veins, c) he untied the string (F_{max} =302 Hz and F_{min} =98 Hz).*

3). Relative constant register position within the utterance, width register variable. Sad speech comparing to speech in depression has more variability in F_0 contour. Fig. 4 shows example of sadness with 4 pitch accents extracted H_L_, HL (falling pitch accents).

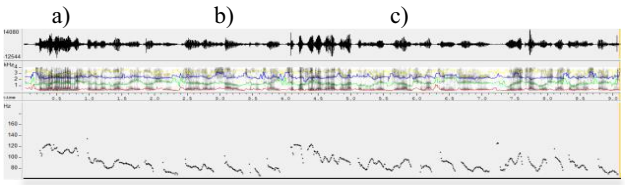


Figure 4: *Pitch accents in the utterances: a) Please, I live on the eighth floor, there is a pub here, b) They constantly go out and scream, c) They are noisy, I can't sleep (F_{max}=122 Hz, F_{min}=78 Hz). Four accents (two H_L and two H_L) were recognized.*

4). Changes in pitch register width and its position

In cases of anger and mixed emotions significant changes of both pitch position and pitch range were observed (**S**). Fig. 5 illustrates F_0 contour for an utterance in a female voice classified as indignation. The speaker can easily control her emotional state so that her message is clearly perceived by the listener. Each syllable which is lexically permissible is clearly stressed. The statistic for the utterances from this group showed all types of accent shapes which were defined for neutral speech, but in 70% with extra dynamic accent, in 26% with shift of register.

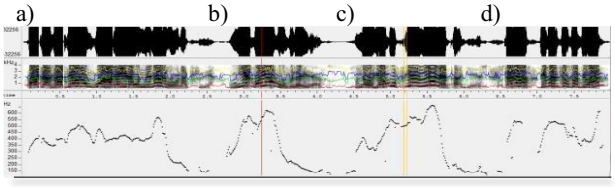


Figure 5: *F_0 contour in 4 phrases from an expressive utterance (indignation): I've got here such a drunkard, he's maltreating me, I am going to thrash him... a) how much he nerves me (F_{max} =569 Hz, F_{min} =149 Hz, F) b) not, because I'm giving him (F_{max} =612 Hz, F_{min} =119 Hz, EH_L) b) I'll fucking thrash him, (F_{max} =659 Hz, F_{min} =133 Hz, EH_L) d) you know what I did (F_{max} =376 Hz, F_{min} =612 Hz, EL_H)*

In the cases of mixed emotions significant changes of both register position and width were observed. Fig. 6 illustrates F_0 contour for an utterance in a female voice (panic).

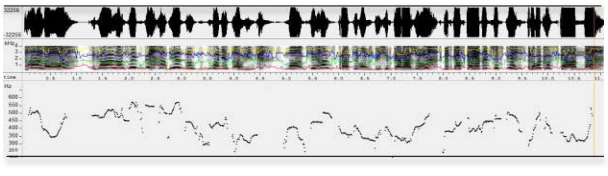


Figure 6: *F_0 contour from an expressive utterance spoken under extreme stress (panic): Hello, I'm calling you. I came from work, my husband has hanged himself in the garage.*

Fig.7 illustrates three pitch accent shapes with extra dynamic accents.

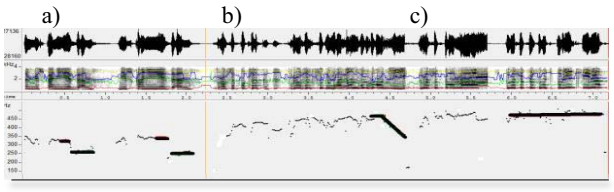


Figure 7: *Pitch accents shapes in 3 expressive phrases: a) Jesus, I call you (F_{max} =570Hz, F_{min} =365Hz, LH_), b) My husband has hanged himself in the garage (F_{max} =549Hz, F_{min} =314Hz, L_H), c) Anna (F_{max} =494Hz, F_{min} =315Hz, L_H).*

Pitch accents statistics have shown that for neutral speech (**N**) the most frequently realized accent is F (40%), common were also the accents for which F_0 change occurred beyond the accented syllable (types: H_H: 36% and L_H_: 28%). For extreme stress (**ES**) the most frequently realized accent is L!H (40%), and HL!, common were also the accents for which F_0 change occurred beyond the accented syllable (types: H_H!: 36%). For limited stress (**S**) the most frequently realized accent is F (40%), common were also the accents for which F_0 change occurred beyond the accented syllable (types: H_H: 36% and L_H_: 28%). For depression (**D**) the most frequently realized accent is F (40%).

5. Discussion

Despite restricting the study to 40 speakers and the limited statistics, a clear tendency in acoustic characterization of expressive speech may be observed.

1) In the current, as well as in a previous study [27], it was stated that a shift in the F_0 contour is an important stress indicator, thus an increase in F_{min} in stressed speech is a result of a shift in the F_0 register, especially when caused by extreme fear.

2) A systematic increase in the register width and position for the stress related to anger or irritation was observed.

3) As the register moves upwards, the accent structures change significantly (there are mainly falling accents, probably related to the tendency of the speaker to minimize the voice effort). Contours become more flat, accentuation is achieved by other prosodic factors.

4) In depression, the F_0 contours are flat, in some cases there is a change in the register (vocal fry).

The results confirmed the need for including such phenomena as the shift of pitch position, change in pitch register width and type of pitch movement into processing expressive utterances, specially produced under stress. Further development of the introduced method demands explanation of the factors that have an influence on pitch register changes in utterances diversified linguistically and in terms of situational context.

6. Acknowledgements

The study was supported by the Polish National Science Centre, project no.: 2014/14/M/HS2/00631, *"Automatic analysis of phonetic convergence in speech technology systems"*.

7. References

[1] Hirst, D. J., "Form and function in the representation of speech prosody", *Speech Communication* 46, 334-347, 2005.

[2] De Looze, C., Hirst, D., "Detecting changes in key and range for the automatic modelling and coding" , In: Speech Prosody, 135-138, 2008.

[3] Patterson, D., Ladd, D. R., "Pitch range modelling: linguistic dimensions of variation", Proc. of ICPhS, 1169-1172, 1992.

[4] Ladd, D. R., "Constraints on the gradient variability of pitch range (or) Pitch level 4 lives!", Phonological Structure and Phonetic Form, 3, 43, 1994.

[5] Hollien, H., "On Vocal Registers", Journal of Phonetics, 2, 125-143, 1972.

[6] Cruttenden, A., "Intonation" Cambridge University Press. Cambridge, England, 1986.

[7] Shriberg, E., Ladd, D.R., Terken, J., Stolcke, A., "Modeling pitch range variation within and across speakers: predicting F0 targets when 'speaking up'.", Proc. of the International Conference on Spoken Language Processing (Addendum, 1- 4), Philadelphia, PA, 1996.

[8] Crystal, D., "Relative and absolute in intonation analysis", J. Int. Phonetic Ass. 1(01), 17-28, 1971.

[9] Hermes, D. J., Van Gestel, J. C., "The frequency scale of speech intonation", J. Acoust. Soc. Am. 90(1), 97-102, 1991.

[10] Honorof, D. N., Whalen, D. "Perception of pitch location within a speaker's F0 range". J. Acoust. Soc. Am. 117(4), 2193-2200, 2005.

[11] Bishop, J., Keating, P., "Perception of pitch location within a speaker's range: Fundamental frequency, voice quality and speaker sex", J. Acoust. Soc. Am. 132(2), 1100-1112, 2012.

[12] O'Shaughnessy, D., Allen, J., "Linguistic modality effects on fundamental frequency in speech", J. Acoust. Soc. Am. 74(4), 1155-1171,1983.

[13] Mozziconacci, S., "Prosody and emotions", Speech Prosody 2002, International Conference, 1-9, 2002.

[14] Hirschberg, J., Pierrehumbert, J., "The intonational structuring of discourse", Proceedings of the 24th Annual Meeting on Association for Computational Linguistics, 136-144, 1986.

[15] Kohler, K., 'Timing and communicative functions of pitch contours", Phonetica 62(2-4), 88, 2005.

[16] Oudeyer, P.-Y., "The production and recognition of emotions in speech: features and algorithms.", Int. J. of Human-Computer Studies 59 (1–2), 157–183, 2003.

[17] Scherer, K. R., "On the nature and function of emotion: A component process approach", Approaches to emotion, 293-317, 1984.

[18] Hansen, J., et al., "The Impact of Speech Under `Stress' on Military Speech Technology." NATO report, http://www-gth.die.upm.es/research/documentation/referencias/Hansen_The Impact.pdf, 2007.

[19] Lefter, J., Rothkrantz, L., Leeuwen, D., Wiggers, P., "Automatic stress detection in emergency (telephone) calls.", International Journal of Intelligent Defence Support Systems 4(2), 148-168 (21), 2011.

[20] Abramson, A. S., "Static and dynamic acoustic cues in distinctive tones". Lang. Speech. 21(4), 319-325, 1978.

[21] Ohala, J. J., Ewan, W. G. 'Speed of pitch change". J. Acoust. Soc. Am. 53(1), 345-345, 1973.

[22] Niebuhr, O., D'Imperio, M., GiliFivela, B., Cangemi, F.,"Are there 'shapers' and 'aligners'? Individual differences in signalling pitch accent category". Proc. 17th ICPhS Hong-Kong, 120-123, 2011.

[23] Demenko G., Oleśkowicz-Popiel M., "Automatic Pitch Accent Annotation", Speech Prosody Proceedings, Boston, 74-78, 2016.

[24] Demenko, G., "Voice stress extraction", Proc. Speech Prosody, Campinas, 53-56, 2008.

[25] Švec, J. G., Schutte, H. K., Miller, D. G.,"On pitch jumps between chest and falsetto registers in voice: Data from living and excised human larynges". J. Acoust. Soc. Am. 106(3), 1523-1531,1991.

[26] Protopapas A., Lieberman P., "Fundamental frequency of phonation and perceived emotional stress.",*J. Acoust. Soc. Am.* 101 (4), 2268-2277, 1997.

The Production of Voicing and Place of Articulation Contrasts by Australian English-Speaking Children

Laurence Bruggeman[1,2], Julien Millasseau[2], Ivan Yuen[2] & Katherine Demuth[2]

[1]The MARCS Institute & ARC Centre of Excellence for the Dynamics of Language, Western Sydney University, [2]Department of Linguistics & ARC Centre of Excellence in Cognition and its Disorders, Macquarie University, Sydney, Australia

Laurence.Bruggeman|Julien.Millasseau|Ivan.Yuen|Katherine.Demuth@mq.edu.au

Abstract

We present an acoustic analysis of voice onset time (VOT) and closure duration (CD) of all six English oral stop consonants in word-initial position, as produced by 4-5- and 9-11-year-old monolingual Australian English-speaking children. Results showed that both groups of children produced a clear voicing contrast at all places of articulation, with longer VOTs and shorter CDs for voiceless than for voiced stops. VOT and CD were inconsistently used by both groups to distinguish between bilabials and non-bilabials but not between alveolars and velars. No significant differences in VOTs and CDs were observed between age groups.

Index Terms: VOT, closure duration, stop consonants, speech production, Australian English, language acquisition, children

1. Introduction

The English phoneme inventory contains six oral stop consonants, which can be classified according to their voicing (voiced: /b, d, g/; voiceless: /p, t, k/) and place of articulation (PoA; bilabial: /b, p/; alveolar: /d, t/; velar: /g, k/). As part of their language development, children acquiring English form phonological categories and develop the ability to accurately produce each of these oral stops. Here, we examine the production of oral stop consonants by Australian English-speaking children, by looking at two temporal cues, namely voice onset time (VOT) and closure duration (CD).

VOT is one of the primary cues used to distinguish voiced from voiceless stops [1], and is defined as the time between the release of the burst and the onset of voicing in the following vowel. English voiceless stops are realised with a long-lag VOT, while English voiced stops are typically produced with short-lag VOTs – although pre-voicing (i.e., onset of voicing *before* the burst release, indicated with negative VOT values) is occasionally reported (e.g., in Australian English [2, 3] and Canadian English [4]). The VOT values of each stop consonant vary depending on the dialect of English that is spoken (e.g., [2-5]). To the best of our knowledge, only [2], [3] and [6] have provided values for Australian English. They are shown in Table 1.

By the time they are one month of age, infants acquiring English as their first language have already started to become perceptually sensitive to the voicing distinction between /b/ and /p/ [7]. It takes children considerably more time, however, to develop the ability to systematically *convey* voicing distinctions in their speech. In English-speaking children, this ability appears to emerge between 2 and 4 years of age, yet the contrast is realised with VOTs that are not yet adult-like and quite variable [8-10]. The variability decreases until it stabilises around age 8-9 [11] and it takes until around age 11 for the actual VOT values to become adult-like [10].

It has been suggested that apart from cueing voicing, VOT may also be used as a cue to contrast for PoA, with VOT increasing as the PoA moves from bilabial to alveolar to velar (i.e., from the front of the mouth to the back) [1, 12, 13]. However, several studies on various dialects of English have failed to fully confirm this pattern (e.g., [5, 14-16]), suggesting that other cues in addition to VOT may be needed to contrast for PoA. One likely cue is CD, which is the duration of the period during which the vocal tract is fully closed to allow air pressure to build up in preparation for the release burst of the stop consonant. While it is undecided whether CD serves as a cue to voicing (e.g., [17, 18]), CD may cue some PoAs in both perception and production [13, 17-19], although this has mainly been investigated in stop consonants that were not word-initial.

Another candidate is the relationship between VOT and CD, as a trade-off between these two temporal cues has previously been suggested, with VOT shortening as CD increases [12]. One may argue this relationship could be especially useful when looking at stops produced by children of various ages, as children's speaking rate increases with age [20] and both VOT and CD are affected by speaking rate [21].

At present, no data on children's production of oral stop consonants is available for Australian English, yet this data is essential for future research, for instance to investigate the acquisition and development of voicing contrasts in children for whom these contrasts are likely to be difficult – such as children acquiring English as a second language, or children with hearing loss. In the present study we thus aimed to investigate the production of all six English oral stop consonants (/b, d, g, p, t, k/) in word-initial position by 4-5 and 9-11-year-old monolingual Australian English-speaking children. More specifically, we aimed to examine how these children use VOT, CD, and the relationship between them to realise voicing and PoA contrasts. This much-needed baseline may then be used in future research.

We predict that both younger and older children will produce a voicing contrast, with shorter VOTs and possibly longer CDs for voiced stops than for voiceless stops. We also predict that they may use VOT and CD – and possibly the relationship between them – as a cue to contrast for some but not all places of articulation.

Table 1. *VOT (ms) for Australian English stops. Values from [2] are approximates.*

study	/b/	/d/	/g/	/p/	/t/	/k/
[2]	-28	5	10	50	70	70
[3]	-2	6	-	77	83	-
[6]	14	26	26	86	101	98

2. Method

2.1. Participants

Twenty monolingual Australian English-speaking children were recruited through the participant register of the Child Language Lab at Macquarie University. Group YC consisted of ten Younger Children (5M, 5F; M_{age} = 5.0, SD = 0.43), while group OC was made up of ten Older Children (5M, 5F; M_{age} = 10.3 years; SD = 0.91). Participants were reimbursed for their participation, and YC participants received a sticker chart.

2.2. Stimuli

Six sets of three monosyllabic English CVC words were created so that all words in a set started with one of the stop consonants /b/, /d/, /g/, /p/, /t/, and /k/, followed once by each of the vowels /ɪ/, /ɐ/, and /ɔ/ (see Table 2 for a full list of target words). Target words were high-frequency words, with a mean frequency of 4.5 Zipf in the Subtlex-UK CBeebies pre-schooler corpus [22]. The 6 (consonants) x 3 (vowels) = 18 target words were embedded in the carrier sentences "See this X" (for nouns) and "These mice X" (for verbs), which were recorded by a female native speaker of Australian English. An additional three sentences were recorded for the practice trials. Each audio recording was paired with a cartoon image of the target word.

Table 2. *Target words used in the experiment.*

	/b/	/d/	/g/	/p/	/t/	/k/
/ɪ/	bib	dig	give	pig	tip	kid
/ɐ/	bug	duck	gut	puff	tub	cup
/ɔ/	bomb	dog	god	pot	top	cob

2.3. Procedure

Participants were seated at a table in a sound-attenuated room on which an iPad 2 and a microphone were placed. The cartoon images and audio stimuli were presented on the iPad using Keynote in combination with a Genelec 8020C speaker. One second after an image was displayed, participants heard the accompanying audio stimulus. Participants were instructed to say the sentence they heard and tap the screen once they had finished speaking to advance to the next trial. After participants had completed three practice trials, they completed five blocks that each contained all 18 target words in a different pseudo-randomised order, so that 18 (words) x 5 (repetitions) = 90 tokens were produced per participant. Responses were recorded with a Shure KSM137 (YC group) and an AKG C535EB (OC group) microphone in Audacity and exported as 44.1 Khz 16-bit WAV files. Children in the YC group received a sticker on a 'treasure hunt chart' after each block to keep them motivated. Informed consent was obtained from all participants and their parents before the start of the experiment.

2.4. Acoustic coding

Acoustic analyses were conducted in Praat. On six occasions, an older child unintentionally tapped the screen twice upon completion of a trial, thereby skipping a trial, so only 894 out of 900 possible tokens were recorded from group OC. There were no missing trials from the younger children as the experimenter was present in the room with them and corrected any accidental misses. Of the recorded tokens, we excluded any tokens that were interrupted by background noise, where participants did not produce the intended target word, or where no release burst of the stop consonant could be observed. Thus, 46 YC tokens (5.1%) and 1 OC token (0.1%) were excluded. In total, 854 YC and 893 OC tokens were acoustically coded for VOT, which was measured from the beginning of the release burst until the onset of a strong F2. VOTs further than 2 SD from the mean were considered outliers and excluded from further analysis. This was the case for 4.0% of tokens in each group. CD was determined for all remaining tokens, as measured from the end of frication of the last segment of the carrier phrase to the beginning of the release burst. Tokens containing a pause or hesitation between the carrier phrase and the target word, as well as CDs further than 2 SD from the mean were discarded from the analyses involving CD. This was the case for 42 YC and 55 OC tokens.

3. Results

3.1. VOT

The mean VOTs from both participant groups are shown in Table 3, while Figure 1 shows the distribution of the VOTs.

Table 3. *VOT in ms (SD) for all six stops as produced by the younger and older children.*

		/b/	/d/	/g/	/p/	/t/	/k/
VOT	younger	10.0 (7.9)	21.7 (12.5)	23.8 (10.1)	99.8 (39.5)	99.6 (39.0)	102.2 (39.2)
	older	14.3 (7.0)	24.3 (9.4)	29.0 (8.9)	80.6 (23.9)	92.0 (24.7)	93.6 (25.7)

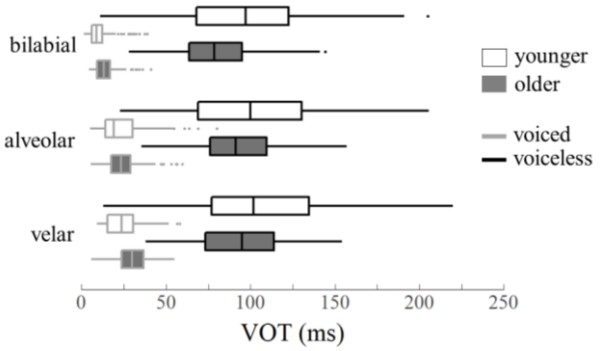


Figure 1. *Distribution of VOT in ms by PoA.*

To investigate whether participants used VOT to contrast for voicing and PoA, VOT was analysed with a linear mixed-effects regression model (LMER) in R, using the packages lme4 [23], lmerTest [24] and lsmeans [25]. Group (YC coded as 1 and OC as -1), Voicing (Voiceless coded as 1 and Voiced as -1), and PoA (Helmert coded; PoA1 compared Bilabial to the mean of Alveolar and Velar, and PoA2 compared Alveolar to Velar) were entered into the model as fixed categorical predictors. Random intercepts for participants and items were added to the model, as well as random slopes for Voicing and PoA by participant. Results of this regression analysis are displayed in Table 4 and show main effects of Voicing and PoA1 (i.e., Bilabial vs the mean of Alveolar and Velar), a 2-way interaction between Voicing and PoA1, and a 3-way interaction between Voicing, PoA1 and Group. Post-hoc analyses using Tukey-adjusted α-levels revealed that both participant groups produced a voicing contrast, with significantly longer VOT for voiceless than voiced stops at each PoA. They also showed that while the OC group produced

Table 4. *Results of the LMER analyses of VOT, of closure duration and of the closure duration-VOT ratio.*

Effect	VOT			Closure duration			CD-VOT ratio		
	Est.	SE	*t*-value	Est.	SE	*t*-value	Est.	SE	*t*-value
(Intercept)	57.64	2.74	21.00*	118.82	5.60	21.23*	0.69	0.01	76.50*
Group	2.11	2.71	0.78	13.56	5.54	2.45*	0.02	0.01	2.55*
Voicing	37.13	2.18	17.05*	-14.16	1.56	-9.09*	-0.17	0.01	-23.43*
PoA1 (bilabial vs non-bilabial)	-9.47	1.48	-6.39*	25.35	3.48	7.28*	0.08	0.01	9.72*
PoA2 (alveolar vs velar)	-2.71	1.57	-1.72	-0.52	4.08	-0.13	0.01	0.01	0.72
Group * Voicing	4.11	2.14	1.93	-0.85	1.33	-0.64	0.00	0.01	-0.42
Group * PoA1 (bilabial vs non-bilabial)	2.36	1.19	1.99	4.49	3.03	1.48	-0.01	0.01	-1.98
Group * PoA2 (alveolar vs velar)	0.10	1.19	0.08	4.27	3.57	1.20	0.01	0.01	1.03
Voicing * PoA1 (bilabial vs non-bilabial)	3.05	1.29	2.36*	-5.00	2.18	-2.29*	-0.01	0.01	-2.23*
Voicing * PoA2 (alveolar vs velar)	0.63	1.50	0.42	-0.16	2.53	-0.06	-0.01	0.01	-1.42
Group * Voicing * PoA1 (bilabial vs non-bilabial)	2.58	0.94	2.74*	-3.83	1.35	-2.83*	-0.01	0.00	-2.18*
Group * Voicing * PoA2 (alveolar vs velar)	-1.06	1.10	-0.96	1.05	1.58	0.67	0.01	0.00	1.63

*$p < .05$

significantly shorter VOTs for the bilabial stops than for the alveolar and velar stops in both voicing conditions (/b/-/d/: $p = .046$; /b/-/g/: $p < .001$; /p/-/t/: $p = .025$; /p/-/k/: $p = .006$) the YC group did this in the voiced (/b/-/d/: $p = .013$; /b/-/g/: $p < .001$) but not the voiceless (/p/-/t/: $p = 1$; /p/-/k/: $p = .995$) condition.

3.2. Closure duration

Table 5 shows the mean CD from both participant groups. To investigate whether CD was used to produce voicing and PoA contrasts, we constructed an LMER model identical to that used in section 3.1 but used CD as the dependent variable instead of VOT. The results of this model are shown in the middle panel of Table 4 and show a main effect of Group, of Voicing, and of PoA1 (i.e., Bilabial vs the mean of Alveolar and Velar), as well as interactions between Voicing and PoA1, and between Group, Voicing and PoA1. Post-hoc analyses showed that both participant groups produced voiceless stops with significantly shorter CD than voiced stops at each PoA, which confirms that they used CD to cue voicing. In addition, the OC group used CD to contrast between bilabials and alveolars in both voicing conditions (/b/-/d/: $p = .012$; /p/-/t/: $p = .020$) but not between any other PoAs. The YC group, on the other hand, used CD to contrast between bilabials and velars in both voicing conditions (/b/-/g/: $p < .001$; /p/-/k/: $p = .034$), and to contrast voiced but not voiceless bilabials and alveolars (/b/-/d/: $p < .001$).

Table 5. *Closure duration in ms (SD) for all six stops as produced by the younger and older children.*

		/b/	/d/	/g/	/p/	/t/	/k/
closure duration	**younger**	171.6 (49.6)	135.4 (46.3)	132.8 (41.9)	131.8 (35.7)	113.7 (41.2)	107.4 (30.2)
	older	133.2 (32.9)	108.3 (38.2)	113.8 (35.4)	104.1 (27.2)	83.9 (27.7)	88.5 (29.3)

3.3. Ratio of closure duration to VOT

To investigate the relationship between VOT and CD, and to account for potential differences in speaking rate between participant groups, CD was divided by the sum of CD and VOT. The means for the resulting ratio are shown in Table 6. Again, an LMER model identical to that used in section 3.1 was constructed, this time using the ratio as the dependent variable. As can be seen in the right panel of Table 4, the model revealed a main effect of Group, of Voicing, and of PoA1 (i.e., Bilabial vs the mean of Alveolar and Velar), as well as interactions between Voicing and PoA1, and between Group, Voicing and PoA1. Post-hoc analyses showed that both participant groups had a significantly lower ratio for voiceless than voiced stops at each PoA. They also showed that, with the exception of the YC group's /p/-/t/ contrast ($p = .238$), both groups' ratios were significantly longer for bilabials than for alveolars and velars in both voicing conditions (OC: /b/-/d/: $p < .001$; /b/-/g/: $p < .001$; /p/-/t/: $p < .001$; /p/-/k/: $p < .001$; YC: /b/-/d/: $p < .001$; /b/-/g/: $p < .001$; /p/-/k/: $p = .028$).

Table 6. *Ratio of CD to VOT (SD) for all six stops as produced by the younger and older children.*

		/b/	/d/	/g/	/p/	/t/	/k/
CD / (CD+VOT)	**younger**	.94 (.05)	.86 (.07)	.84 (.07)	.58 (.11)	.54 (.12)	.53 (.11)
	older	.90 (.05)	.80 (.09)	.79 (.07)	.57 (.07)	.47 (.09)	.48 (.09)

4. Discussion

The present study investigated the production of English oral stops in word-initial position by monolingual Australian English-speaking children. We predicted that both the younger and older children would produce a voicing contrast, and that they would use VOT and potentially also CD and the CD-VOT ratio to realise this contrast. This prediction was borne out, as both participant groups produced significantly longer VOTs and shorter CDs (thus resulting in lower CD-VOT ratios) for voiceless than voiced stops and did so at each PoA.

We further predicted that the children may use VOT and CD as a cue to contrast for some but not all PoAs, which was also confirmed. The only PoA contrasts that were made using any of the examined temporal cues were between bilabials and non-bilabials. Neither group used VOT, CD or the CD-VOT ratio to distinguish alveolar from velar stops. This is in line with the pattern that has previously been found for Australian English-speaking adults by [2] and [6] and for adult speakers of other dialects of English [5, 13, 15]. While the OC group used VOT *and* CD to produce a contrast between bilabials and alveolars, and VOT *but not* CD to contrast bilabials and velars, the YC group showed a different pattern. They used VOT *and* CD for the distinction between bilabials and both velars and alveolars,

but only for the voiced stops. For the voiceless stops, the only PoA contrast realised by the younger children was that between bilabials and velars, using CD *but not* VOT. Neither VOT, CD nor the CD-VOT ratio was used by the younger children to contrast between voiceless bilabials and alveolars (i.e., /p/-/t/). The fact that the YC group contained only ten children, whose acoustic cues were quite variable due to their young age, may have prevented us from finding this latter PoA contrast.

No significant differences were found between the younger and older children's VOTs, CDs or ratios at any of the PoAs. However, overall, the YC appeared to be more variable on all three dependent variables than the OC (although we did not test this statistically). This would be in line with findings from [11], who found that children's VOT variability decreases until it stabilises at age 8-9.

While a direct statistical comparison with adult VOT data for Australian English is not possible, the present data may nonetheless be compared indirectly to the available values shown in Table 1. The VOTs from the present study seem quite different from those found by [2] and, to a lesser extent, [3]. This is likely due to geographical and methodological differences, as the data from [2] were obtained from four speakers from the Northern Territory by means of a single-word picture naming task, while [3] used syllables embedded in a carrier sentence that speakers read out from a computer screen. Indeed, our VOT values appear quite similar to – albeit a little more variable than – those reported for adults by [6], who recorded speakers from the same geographical area (Sydney, New South Wales) and used a similar experimental task (elicited imitation using carrier phrases).

In sum, this study has provided VOT and CD data for oral stop consonants produced by children acquiring Australian English. It thus contributes much-needed baseline data that have so far not been available for this dialect of English. We recommend that future studies investigating the production of stop consonants in Australian English record data from a larger sample of children in each age group and aim to include a control group of adult speakers so that the adult-likeness of the child productions may be compared directly.

5. Acknowledgements

This work was funded by the ARC Laureate Fellowship [FL130100014] awarded to the last author, and by the ARC Centre of Excellence in Cognition and its Disorders [CE110001021]. We thank Phillip (Xin) Cheng for his assistance with the acoustic coding.

6. References

[1] Lisker, L., and Abramson, A.S., "A cross-language study of voicing in initial stops: Acoustical measurements", WORD, 20(3): 384-422, 1964.

[2] Jones, C., and Meakins, F., "Variation in voice onset time in stops in Gurindji Kriol: Picture naming and conversational speech", Aust J Linguist, 33(2): 196-220, 2013.

[3] Antoniou, M., Best, C.T., Tyler, M.D., and Kroos, C., "Language context elicits native-like stop voicing in early bilinguals' productions in both L1 and L2", J Phon, 38(4): 640-653, 2010.

[4] Caramazza, A., Yeni- Komshian, G.H., Zurif, E.B., and Carbone, E., "The acquisition of a new phonological contrast: The case of stop consonants in French- English bilinguals", J Acoust Soc Am, 54(2): 421-428, 1973.

[5] Docherty, G.J., The timing of voicing in British English obstruents, Foris Publications, 1992.

[6] Millasseau, J.: 'The acquisition of voicing contrasts in Australian English-speaking 4-year-olds'. Master's Thesis, Macquarie University, 2017.

[7] Eimas, P.D., Siqueland, E.R., Jusczyk, P., and Vigorito, J., "Speech perception in infants", Science, 171(3968): 303-306, 1971.

[8] Kewley-Port, D., and Preston, M.S., "Early apical stop production: A voice onset time analysis", J Phon, 2(3): 195-210, 1974.

[9] Barton, D., and Macken, M.A., "An instrumental analysis of the voicing contrast in word-initial stops in the speech of four-year-old English-speaking children", Lang Speech, 23(2): 159-169, 1980.

[10] Yu, V.Y., De Nil, L.F., and Pang, E.W., "Effects of age, sex and syllable number on voice onset time: Evidence from children's voiceless aspirated stops", Lang Speech, 58(2): 152-167, 2015.

[11] Eguchi, S., and Hirsch, I.J., "Development of speech sounds in children", Acta Otolaryngol, 68(sup257): 1-51, 1969.

[12] Cho, T., and Ladefoged, P., "Variation and universals in VOT: evidence from 18 languages", J Phon, 27(2): 207-229, 1999.

[13] Yao, Y., "Closure duration and VOT of word-initial voiceless plosives in English in spontaneous connected speech", UC Berkeley Phonology Lab Annual Reports, 3(3), 2007.

[14] Yao, Y., "Understanding VOT Variation in Spontaneous Speech", UC Berkeley Phonology Lab Annual Reports, 5: 29-43, 2009.

[15] Stuart-Smith, J., Sonderegger, M., Rathcke, T., and Macdonald, R., "The private life of stops: VOT in a real-time corpus of spontaneous Glaswegian", Lab Phonol, 6(3-4): 505, 2015.

[16] Nearey, T.M., and Rochet, B.L., "Effects of place of articulation and vowel context on VOT production and perception for French and English stops", J Int Phon Assoc, 24(1): 1-18, 1994.

[17] Port, R.F., "The influence of tempo on stop closure duration as a cue for voicing and place", J Phon, 7(1): 45-56, 1979.

[18] Luce, P.A., and Charles- Luce, J., "Contextual effects on vowel duration, closure duration, and the consonant/vowel ratio in speech production", J Acoust Soc Am, 78(6): 1949-1957, 1985.

[19] Repp, B.H., "Closure duration and release burst amplitude cues to stop consonant manner and place of articulation", Lang Speech, 27(3): 245-254, 1984.

[20] Kowal, S., O'Connell, D.C., and Sabin, E.J., "Development of temporal patterning and vocal hesitations in spontaneous narratives", J Psycholinguist Res, 4(3): 195-207, 1975.

[21] Miller, J.L., Green, K.P., and Reeves, A., "Speaking rate and segments: A look at the relation between speech production and speech perception for the voicing contrast", Phonetica, 43(1-3): 106-115, 1986.

[22] Van Heuven, W., Mandera, P., Keuleers, E., and Brysbaert, M., "Subtlex-UK: A new and improved word frequency database for British English", 67: 1176-1190, 2014.

[23] Bates, D., Maechler, M., Bolker, B., and Walker, S., "Fitting linear mixed-effects models using lme4", J Stat Softw, 67(1): 1-48, 2015.

[24] 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.

[25] Lenth, R.V., "Least-Squares Means: The R Package lsmeans", J Stat Softw, 69(1): 1-33, 2016.

Investigation of DNN Prediction of Power Spectral Envelopes for Speech Coding & ASR

Christine Pickersgill[1], *Stephen So*[1,2], *Belinda Schwerin*[1,2]

[1] School of Engineering and Built Environment, Griffith University, Australia
[2] Signal Processing Laboratory, Griffith University, Australia

christine.pickersgill@griffithuni.edu.au; s.so@griffith.edu.au; b.schwerin@griffith.edu.au

Abstract

This paper proposes a DNN-based preprocessing method for speech coding and automatic speech recognition applications. The method proposed here maps noisy log power spectra to "clean" smoothed log power spectral envelopes using DNN prediction. The proposed method has the advantage of combining feature extraction with DNN-based enhancement, thus reducing computational time and resources. The TIMIT speech database with various additive noise types was used to train the DNN, and the NN prediction results are compared to the target clean log power spectral envelopes using log spectral distortion. The proposed method is found to have lower log spectral distortion measurements compared to similar neural networks that map noisy power spectra to clean power spectra.

Index Terms: Deep neural networks, speech coding, preprocessing, automatic speech recognition, power spectra.

1. Introduction

This paper investigates the use of a neural network (NN) as a preprocessing method for speech coding and automatic speech recognition (ASR). Various types of noise are often present in recorded speech, which can have a degrading effect on speech coding and ASR algorithms.

Many approaches to reducing the effects of noise on ASR have been investigated over the years. One approach is to use speech enhancement as a preprocessor, prior to extracting any features necessary for coding or recognition [1]. Another approach is to extract more robust features to begin with, such as Mel Frequency Cepstral Coefficients (MFCCs) or Perceptual Linear Prediction (PLP) coefficients, which exploit the human ear's perceptual properties to reduce noise interference [2][3]. Yet another approach to increasing the robustness of ASR is known as feature enhancement, which subjects the noisy features to some kind of processing after extraction but before recognition [4][5].

Many of the above techniques involve framing, processing, and reconstructing noisy speech files into enhanced intelligible speech, before processing again to extract features for speech coding or ASR. If the noisy speech only needs to be used for coding or recognition purposes, however, then it does not actually need to be reconstructed fully – it is only necessary to estimate parameters that represent the power spectrum of the speech, such as Linear Prediction Coefficients (LPCs). If it is possible to achieve an estimate of clean power spectra when presented with noisy speech, these predicted clean power spectra parameters could be directly utilised by the speech coding or recognition system.

Previous work has investigated the use of a deep neural network (DNN) to map noisy log power spectra to clean log power spectra as a speech enhancement method, resulting in improved quality and intelligibility in the reconstructed speech [6][7]. Speech coding and ASR, however, only require the smoothed power spectral envelope. For this reason, the method proposed here is the mapping of noisy log power spectra to clean smoothed log power spectral envelopes using a deep neural network (DNN).

As the smoothed log spectral envelope contains less fine detail than the log power spectra, the proposed method offers less computational and numerical complexity – essentially an "easier" prediction for the DNN – with the added advantage of combining feature extraction with DNN-based enhancement.

The rest of this paper is organised as follows: Section 2.1 outlines the feature extraction method used to create the NN training set; the NN architecture is described in Section 2.2; the speech database and evaluation method are discussed in Section 3; and experimental results are shown in Section 4.

2. Method

This paper proposes training a NN to map noisy log power spectra to clean smoothed log power spectral envelopes. As shown in fig. 1, the training stage begins with noisy and clean speech samples from a speech database training set, from which log power spectra and smoothed log power envelope features are calculated. The NN is then trained using noisy log power spectra as input and clean smoothed log power envelopes as the target output, resulting in a NN model. In the testing stage the NN model is used to predict clean smoothed log power envelopes when given new noisy log power spectra extracted from the test set. Linear Prediction Coefficients (LPCs) are then derived from the NN prediction.

2.1. Feature extraction

To begin with, the time domain speech samples are segmented into frames of length 25ms with 5ms overlap, followed by the application of a hamming window and then a pre-emphasis filter ($\alpha = 0.97$) to remove the spectral tilt. The autocorrelation of each frame is then found,

$$R_{xx}(k) = \frac{1}{N} \sum_{n=0}^{N\text{-}1\text{-}k} x(n)x(n+k), \qquad (1)$$

where N is the frame length in samples, k is the sample lag, and $x(n)$ is the signal. The next step is to estimate the parameters of an autoregressive (AR) model [8], whose system function is

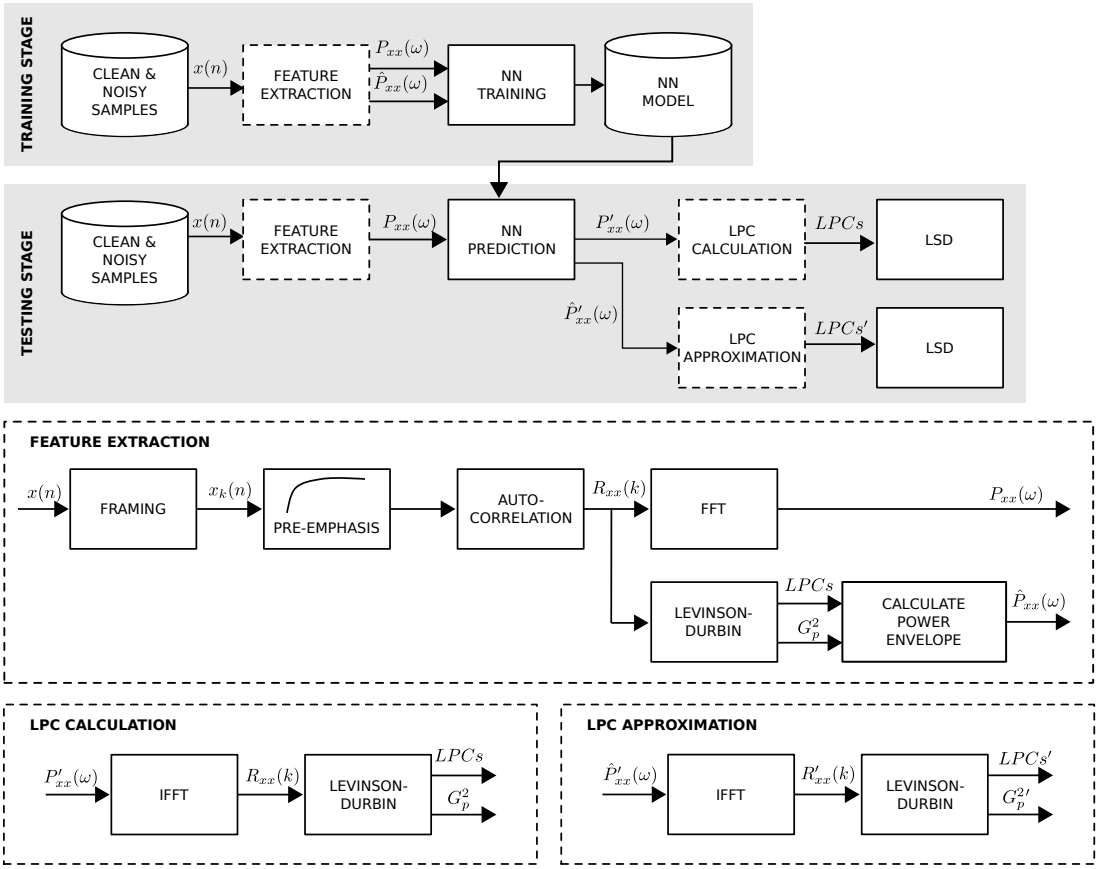


Figure 1: *Block diagram of the proposed NN-based preprocessing method, encompassing feature extraction, NN training, NN prediction, and evaluation of results.*

given by:

$$H(z) = \frac{G_p}{1 + \sum_{k=1}^{p} a_{pk} z^{-k}} \tag{2}$$

Applying the Levinson-Durbin algorithm [9],the a_{pk} values are determined. Thus we find the AR coefficients (or LPCs) of each frame, and the excitation variance G_p^2.

In order to get the feature vectors to train the neural network, it is necessary to get an estimate of the log power spectrum, which will then be used as the input to the NN, and the smoothed log power spectral envelope will then function as the target output. For this application the periodogram $P_{xx}(\omega)$ is used for the estimate of the power spectrum [9],

$$P_{xx}(\omega) = \frac{1}{N} \left| \sum_{n=0}^{N-1} x(n) e^{-j\omega n} \right|^2, \tag{3}$$

where N is the frame size. The LPC values calculated earlier are used to find the smoothed power spectral envelope $\hat{P}_{xx}(\omega)$,

$$\hat{P}_{xx}(\omega) = \frac{G_p^2}{\left| \sum_{k=0}^{N-1} a_{pk} e^{-j\omega k} \right|^2}, \tag{4}$$

where $a_{p0} = 1$. The power spectra and smoothed envelopes are then converted into the log domain.

After training on these features, the NN prediction delivers estimated smoothed log power spectral envelope parameters. From these predicted features it is possible to first apply a log to linear conversion to each parameter vector, then perform an inverse FFT to get the autocorrelation coefficients, and use the Levinson-Durbin algorithm to estimate the LPCs and excitation variance.

2.2. Neural network architecture

The NN architecture and naming scheme are outlined briefly in Table 1 and Table 2, based on a similar architecture arrangement proposed in previous works [6][7], which have been simulated here for the sake of comparison. In each case represented in Table 1 the depth (number of hidden layers) and width (number of nodes per layer) varied, but all other parameters remained constant. In each case represented in Table 2, the depth and width were the constant elements. All NNs used either Stochastic Gradient Descent (SGD) or Adam [10] as an optimiser with a learning rate of 1×10^{-5}. Type C shown in Table 2 is the proposed method.

The model type used was 'sequential', the layer type was 'dense', and the initialisation for each layer was 'uniform'. The activation for each hidden layer was either sigmoid (Type A) or rectified linear [11][12][13]('ReLU', Types B & C), and the activation for the output layer was 'linear'. Input and output layers had 1539 nodes (Type A) or 513 nodes each (Types B & C). The model was fitted using 50 epochs with a mini-batch size

Table 2: *Summary of neural network architecture naming scheme, showing input dimension, number of epochs, learning rate, optimiser, hidden layer activation, input features, and target output (power spectra or smoothed envelope). Type C is the proposed method.*

NN Type	I/P Dim.	No. Epochs	Learning Rate	Opt.	Activation	I/P Features	Target O/P
Type A	1539	50	1×10^{-5}	SGD	Sigmoid	$P_{xx}(\omega)$	$P_{xx}(\omega)$
Type B	513	50	1×10^{-5}	Adam	ReLU	$P_{xx}(\omega)$	$P_{xx}(\omega)$
Type C	513	50	1×10^{-5}	Adam	ReLU	$P_{xx}(\omega)$	$\hat{P}_{xx}(\omega)$

Table 1: *Summary of neural network architecture naming scheme, showing number and size of hidden layers (HL).*

NN Type	No. HL	Size HL
SNN	1	6144
DNN_1	1	2048
DNN_2	2	2048
DNN_3	3	2048

of 128, and mean squared error (MSE) was used as a cost function metric. It is worth noting that in this particular case MSE is equivalent to log spectral distortion (LSD) [14] due to the use of log power spectra as feature vectors. Later on LSD is utilised as an evaluation metric for the NN predicted features, since it is frequently used in speech coding and ASR evaluations.

3. Speech corpus & evaluation

The experiments presented in this paper were done using the TIMIT speech database [15], which contains 6300 utterances by 630 speakers, both male and female, from various dialect regions in the United States. The database was sampled at 16 kHz and has been divided into a training set and a test set. The entire speech corpus was processed to include six types of additive noise from the Noisex-92 noise database (babble, F16, factory, pink, volvo, and white)[16] at various levels of SNR (0dB, 5dB, 10dB, 15dB, 20dB). This created a dataset of several hundred hours of multi-condition training and testing data. The complete training set was used for NN training, consisting of 4620 utterances in their original clean state, with the addition of the noise types and SNRs mentioned above, for a total of 194,040 utterances, or 88 million training vectors.

The complete test set (1680 utterances) was used for testing and prediction, consisting of the original clean utterances in addition to the six noise types mentioned above, though only at SNR levels of 10dB, 15dB, and 20dB. This amounted to a total of 35,280 test utterances, or 18.7 million test vectors.

Log spectral distortion (LSD) [14] was chosen as the evaluation method for this project due to its frequency of use in speech coding and ASR evaluations. LSD has been shown to be equivalent to the squared error of two cepstra, and since the cepstrum is the feature set used in speech recognition, this allows LSD to be interpreted as an indicator of ASR performance [17]. With regards to speech coding applications, past investigations have found that LSD is a good indicator of speech quality [18].

4. Results

A summary of LSD measurements for a selection of noise types and input SNRs is given in Table 3. The noise type and SNR reflects the additive noise with which the original DNN input features were corrupted. All output predicted vectors were compared to unprocessed clean smoothed power spectral envelopes, giving LSD values in dB. With the exception of the *clean* column (in which Type B-DNN3 has the lowest LSD), Types C-DNN2 and C-DNN3 contain the LSD values that are the lowest in their respective columns, indicating the improved quality of the DNN predictions from the proposed method. Note that many of the LSD values in the *clean* column are higher than those in the subsequent noisy columns, which is a result of training the NN on both clean and noisy vectors so as to not over-fit the model to the clean or 'ideal' training situation.

LSD values for DNN Type A are shown in the first segment of Table 3. Type A was based on the method proposed by Xu et al. [6][7], and is shown here for the purpose of comparison with the proposed method. Three feature vectors were concatenated to create input vectors of width 1539, SGD was the learning rate optimiser, sigmoid was the activation function, and the target output was clean power spectra. The LSD values for Type A are the highest in the table, indicating the poor quality of the DNN-predicted vectors.

Type B represents a halfway point between previous works (represented by Type A) and the proposed method, Type C. DNN Type B still maps noisy power spectra to clean power spectra, as in Type A, but it uses single input vectors of width 513 and the same optimiser (Adam) and activation function (ReLU) as Type C. The LSD values for Type B demonstrate the improvement gained by simply changing the optimiser and activation function, supporting the general consensus that Adam and ReLU improve DNN performance [10][11][12][13].

The proposed method is shown as Type C, in which Adam was the optimiser and ReLU was the activation function, as in Type B, but this time input power spectra were mapped to target clean smoothed power spectral envelopes. Most LSD values are lowest for Type C, indicating the higher quality of the DNN prediction, and suggesting that an improvement in ASR rates may be gained through the use of this approach. Comparing Type C LSD values to those of Type B also shows that it is not just the change in optimiser and activation function that improves the DNN prediction, but mapping to smoothed power envelopes increases the DNN's performance yet again.

5. Conclusions

The method proposed here suggests the use of a DNN to map noisy log power spectra to clean smoothed log power spectral envelopes, as an alternative preprocessing method for speech coding or ASR applications. Since the smoothed power enve-

Table 3: *Log Spectral Distortion scores for NN predictions of smoothed power envelopes, for various input noise types and SNRs.*

NN Type	Noise Type	Mean LSD (dB)									
		Clean	Volvo			Babble			Factory		
	SNR (dB)	-	10	15	20	10	15	20	10	15	20
Type A											
A-SNN		6.09	6.20	5.99	5.96	7.75	7.17	6.83	8.37	7.52	7.01
A-DNN1		8.14	8.03	7.78	7.80	9.54	9.04	8.73	9.55	8.97	8.66
A-DNN2		7.28	6.61	6.61	6.67	7.76	7.40	7.23	8.03	7.45	7.17
A-DNN3		8.49	8.20	8.12	8.14	9.35	8.96	8.80	9.85	9.54	9.41
Type B											
B-SNN		5.82	3.48	3.64	3.94	6.18	5.19	4.66	6.73	5.51	4.84
B-DNN1		6.58	3.48	3.55	3.73	5.85	5.03	4.58	6.46	5.36	4.72
B-DNN2		7.45	3.69	3.92	4.26	5.81	5.17	4.84	6.34	5.42	4.94
B-DNN3		**4.33**	3.44	3.72	3.99	5.81	5.01	4.69	6.71	5.56	4.96
Type C											
C-SNN		5.11	3.64	3.68	3.85	6.30	5.45	4.85	6.72	5.68	4.94
C-DNN1		8.65	8.58	7.28	6.64	6.42	6.03	5.86	6.82	6.01	5.63
C-DNN2		5.91	**3.01**	**2.96**	**3.14**	5.85	4.87	**4.27**	6.06	5.00	**4.32**
C-DNN3		5.20	4.07	4.14	4.23	**5.52**	**4.70**	4.43	**5.90**	**4.87**	4.37

lope contains less fine detail than the power spectra, the proposed method offers the advantage of a less numerically complex computation for the DNN.

The proposed method was compared to two other DNN methods that map noisy power spectra to clean power spectra, and the resulting predictions were all measured using LSD. The proposed method performed well when compared to the other two, as evidenced by an improvement in LSD values for many of the noise types and SNRs used for testing. This would seem to suggest that the proposed method shows promise when it comes to feature enhancement for speech coding and ASR.

6. References

[1] K. K. Paliwal, J. G. Lyons, S. So, A. P. Stark, and K. K. Wójcicki, "Comparative evaluation of speech enhancement methods for robust automatic speech recognition," in *Signal Processing and Communication Systems (ICSPCS), 2010 4th International Conference on*. IEEE, 2010, pp. 1–5.

[2] H. Hermansky, "Perceptual linear predictive (plp) analysis of speech," *the Journal of the Acoustical Society of America*, vol. 87, no. 4, pp. 1738–1752, 1990.

[3] S. B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," in *Readings in speech recognition*. Elsevier, 1990, pp. 65–74.

[4] H. Hermansky and N. Morgan, "Rasta processing of speech," *IEEE transactions on speech and audio processing*, vol. 2, no. 4, pp. 578–589, 1994.

[5] B. S. Atal, "Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification," *the Journal of the Acoustical Society of America*, vol. 55, no. 6, pp. 1304–1312, 1974.

[6] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "An experimental study on speech enhancement based on deep neural networks," *IEEE Signal Processing Letters*, vol. 21, no. 1, pp. 65–68, 2014.

[7] ——, "A regression approach to speech enhancement based on deep neural networks," *Audio, Speech, and Language Processing, IEEE/ACM Transactions on*, vol. 23, no. 1, pp. 7–19, 2015.

[8] J. Makhoul, "Linear prediction: A tutorial review," *Proceedings of the IEEE*, vol. 63, no. 4, pp. 561–580, 1975.

[9] M. H. Hayes, *Statistical digital signal processing and modeling*. John Wiley & Sons, 2009.

[10] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[11] G. E. Dahl, T. N. Sainath, and G. E. Hinton, "Improving deep neural networks for lvcsr using rectified linear units and dropout," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 8609–8613.

[12] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *Proceedings of the 27th international conference on machine learning (ICML-10)*, 2010, pp. 807–814.

[13] M. D. Zeiler, M. Ranzato, R. Monga, M. Mao, K. Yang, Q. V. Le, P. Nguyen, A. Senior, V. Vanhoucke, J. Dean *et al.*, "On rectified linear units for speech processing," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 3517–3521.

[14] A. Gray and J. Markel, "Distance measures for speech processing," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 24, no. 5, pp. 380–391, 1976.

[15] J. S. Garofolo *et al.*, "Getting started with the darpa timit cd-rom: An acoustic phonetic continuous speech database," *National Institute of Standards and Technology (NIST), Gaithersburgh, MD*, vol. 107, p. 16, 1988.

[16] A. Varga and H. J. Steeneken, "Assessment for automatic speech recognition: Ii. noisex-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech communication*, vol. 12, no. 3, pp. 247–251, 1993.

[17] L. R. Rabiner and B.-H. Juang, *Fundamentals of speech recognition*. PTR Prentice Hall Englewood Cliffs, 1993, vol. 14.

[18] K. K. Paliwal and B. S. Atal, "Efficient vector quantization of lpc parameters at 24 bits/frame," *IEEE transactions on speech and audio processing*, vol. 1, no. 1, pp. 3–14, 1993.

Author Index