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

6–9 December 2016 ▪ Parramatta, Australia

Conference Information

Table of Contents

Author Index

Search

Editors: Christopher Carignan and Michael D. Tyler

ASSTA

The Australasian Speech
Science & Technology
Association Incorporated

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

Parramatta River Walk, image copyright: Mark Soon Photography

Welcome to Delegates

On behalf of the organising committee, and the Australasian Speech Science and Technology Association (ASSTA), we welcome you to the city of Parramatta for the 16th Australasian International Conference on Speech Science and Technology (SST2016).

Thirty years have now passed since the first SST was held in Canberra in 1986. The first SST was pioneering, as it was the first in the world to combine papers from engineers, phoneticians, computer scientists, speech scientists and psychologists. We have chosen to mark the occasion at this year's conference by inviting the inaugural ASSTA Executive of 1988 to present a special plenary session: Emeritus Fellow Bruce Millar, Emeritus Professor John Clark, Emeritus Professor Michael Wagner, Professor Peter Blamey, and Dr John Ingram. The plenary will celebrate the history of SST, provide insights into the creation of ASSTA, and will highlight the important contribution that speech science and technology in Australasia has made to scientific and technological advances worldwide. We hope that you will enjoy the opportunity to reflect on the history of SST with some of the scientists who were there at the beginning.

For SST2016, 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 in the conference programme). The submissions were all blind peer reviewed by two anonymous reviewers, 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 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. A quarter of the delegates in 2016 are from overseas, hailing from Canada, Germany, Hong Kong, Japan, Malaysia, Saudi Arabia, Singapore, Sweden, Thailand, UK, and USA. We are delighted to welcome our SST2016 international keynotes Harald Baayen, Shri Narayanan, and Leher Singh and we thank them for contributing to the continuing success of SST. Our thanks also go to Aunty Sandra Lee for delivering the Welcome to Country, our special plenary session presenters, and Jonathan Tapson, Director of the MARCS Institute for Brain, Behaviour and Development, for opening the conference.

We would like to acknowledge Western Sydney University for its support of the conference. The conference host, the MARCS Institute for Brain, Behaviour and Development, provided infrastructure, personnel, and financial support, and the School of Humanities and Communication Arts and the School of Social Sciences and Psychology are two of our platinum sponsors. We are very grateful to the ARC Centre of Excellence for the Dynamics of Language for its platinum sponsorship, to the ARC Centre of Excellence in Cognition and its Disorders, Northern Digital Inc., and Cochlear Ltd., for silver sponsorship, and to Objective Eye Tracking, the HEARing Cooperative Research Centre, Karger Medical and Scientific Publishers, and Advanced Bionics, for their bronze sponsorships.

In keeping with SST tradition, this conference has papers from a range of topics across speech science and technology. Please consider venturing into a session on an unfamiliar topic to take advantage of what a multidisciplinary conference has to offer. Note also that there will be a satellite event directly after the conference, supported by ASSTA, on the role of predictability in shaping human language sound patterns.

We hope you enjoy the conference, and that you will also find some time to explore Parramatta, a city with a long history, both before and after European settlement.

Michael Tyler

Paola Escudero

Conference Chairs

ASSTA Corporate Members

appen™

appen.com

Cochlear®

www.cochlear.com

THE HEARING CRC

www.hearingcrc.org

SST2016 Sponsors

Platinum

ARC Centre of Excellence for the Dynamics of Language
MARCS Institute for Brain, Behaviour and Development, Western Sydney University
School of Humanities and Communication Arts, Western Sydney University
School of Social Sciences and Psychology, Western Sydney University

Silver

ARC Centre of Excellence in Cognition and its Disorders
Cochlear Ltd.
Northern Digital Inc.

Bronze

Advanced Bionics
The HEARing Cooperative Research Centre
Karger Scientific Publishers
Objective Eye Tracking

Media Partner

Languages - an open access journal by MDPI

Conference Organisation

Conference Organising Committee

Conference Chairs: Michael Tyler and Paola Escudero

Secretary: Mark Antoniou

Treasurer: Karen Mattock

Technical Program Committee Chair: Chris Carignan

Venue, Accommodation, and Sponsorship: Marina Kalashnikova and Heather Kember

Publicity: Karen Mulak

ASSTA Liaison: Denis Burnham

Conference Secretariat

Craig Purcell

Farah Abdurahman

Andrea Kantek

Christian Roach

Volunteers

Julie Beadle

Jaydene Elvin

Mona Faris

Saya Kawase

Mark Lathouwers

Valeria Peretokina

Stacey Sherwood

Conference Award Winners

ASSTA New Researcher Award

Jia Hoong Ong

Valeria Peretokina

Sarah Wright

CoEDL Early Career Researcher Award

Liquan Liu

Hywel Stoakes

CoEDL Student Award

Rosey Billington

Arwen Blackwood Ximenes

Ann-Kathrin Grohe

Katharina Zahner

HEARing CRC Student Award

Saya Kawase

Objective Eyetracking Student Award

Ben Davies

Nicole Traynor

Review Panel

Waleed Abdulla

Brett Baker

Titia Benders

Tessa Bent

Catherine Best

Jason Brown

Rikke Bundgaard-Nielsen

Ann Burchfield

Felicity Cox

Karen Croot

Chris Davis

David Dean

Katherine Demuth

Donald Derrick

San Duanmu

Ewald Enzinger

Julien Epps

Christopher Fennell

Paul Foulkes

David Grayden

Bernard Guillemin

John Hajek

Jonathan Harrington

Mark Harvey

Rachel Hayes-Harb

Yusuke Hioka

Vincent Hughes

Caroline Jones

Benjawan Kasisopa

Nenagh Kemp

Jeesun Kim

Trent Lewis

Liquan Liu

Debbie Loakes

Robert Mayr

Kiri Mealings

Hansjörg Mixdorff

Geoffrey Morrison

Lyndsey Nickels

Jessie Nixon

Pascal Perrier

Linda Polka

Michael Proctor

Eva Reinisch

Michael Robb

Iris-Corinna Schwarz

Belinda Schwerin

Chilin Shih

Mary Stevens

Anita Szakay

Marija Tabain

William Thorpe

Roberto Togneri

Kimiko Tsukada

Chiharu Tsurutani

Jan van Doorn

Adam Vogel

Catherine Watson

Daniel Williams

Joe Wolfe

Maria Wolters

Ivan Yuen

Cuiling Zhang

Sixteenth Australasian International Conference on Speech Science and Technology

SST 2016

Table of Contents

L1 Acquisition

Speech Prosody I

Australasian Languages

L2 Acquisition I

Applications of Speech Science and Technology

Sociophonetics

Speech Perception I

Special Session: Effects of Early Language Environment and Linguistic Input on the Development of Speech Perception Skills in the First Years of Life

Articulatory Phonetics

Bilingualism

Forensic Speech Science I

Speech Registers

Speech Perception II

Speech Enhancement

Pitch Accent

Morphophonology

Speech Synthesis

Lexical Tone

Infant-Directed Speech

Automatic Speech Recognition

Acoustic Phonetics

Speech Emotion Recognition

Cross-Language Vowel Perception

Forensic Speech Science II

Four-Page Papers Accepted for Poster Presentation

Fundamental frequency characteristics of infant vocalisations: a study in voice quality

Adele Gregory [1], *Marija Tabain* [2]

[1] School of Culture, History and Language, Australian National University, Canberra, Australia
[2] School of Humanities, La Trobe University, Melbourne, Australia
adele.gregory@anu.edu.au, m.tabain@latrobe.edu.au

Abstract

No clear picture exists of the f_0 developmental pattern of typically developing infants. Methodological differences (e.g. type of vocalisations included for analysis) have been found to contribute to this. This paper approaches the f_0 characteristics of infant vocalisations from the perspective of modal and non-modal voice qualities to more fully understand their role in the overall developmental contour. The results presented in this paper support the notion that the f_0 of infant vocalisations provides insight into how an infant learns to exercise vocal control and that voice quality is a useful category through which to investigate these developments.

Index Terms: fundamental frequency, infant, developmental pattern, voice quality

1. Introduction

Previous work conducted on the developmental pattern of typically developing infants' f_0 has been, at times, contradictory. [1], [2] and [3] found no changes in mean f_0 whilst [4] noted a decrease in mean f_0. [3, 918] produced a summary of the available research and concluded (tentatively) that typically developing infants have "high and variable fundamental frequencies in the first year of life in comparison with adults." However no other general trends could be established due to a number of methodological differences in the studies, including:

1. Age of participant,
2. Selection of vocalisations for analysis (e.g. sounds classified as aperiodic, vegetative, squealing, growling),
3. Portion of segment used for f_0 extraction (e.g. nuclei, or nuclei and margins), and
4. How measurements were taken (e.g. visual inspection of waveforms, automatic pitch extraction).

These methodological differences have prevented results being directly compared. Thus a clear understanding of a typically developing infants' f_0 characteristics is yet to be ascertained. This current paper will examine the f_0 characteristics of infant vocalisations by focusing on how the selection of vocalisations for analysis (point 2. above) may give additional insight into a typically developing infant's f_0 characteristics.

It is recognised that vocalisations produced with non-modal voice qualities occupy a large proportion of the sounds an infant produces [5]. These productions include vocalisations with harsh or creaky phonation, those produced with formant structures that are unstable or those with widely fluctuating fundamental frequency. They can also include those that are produced with intermittent voicing or those deemed as vegetative (such as wheezes, sneezes, coughs, hiccups and clicks). Tokens with non-modal voice quality are frequently discarded from analysis in infant developmental studies. In [3, 932] "nearly half of the data had to be discarded because of aperiodicity of the signal, either because syllables were entirely aperiodic (e.g. voiceless sounds) or because they failed to satisfy [their] criterion of having 80% or greater measurable f_0 intervals." Whilst this preference to examine vocalisations with normal phonation is understandable due to it being indicative of emerging linguistic control there is a growing awareness that both modal and non-modal voice uses are important in infants' development of vocal control [5, 553]. This study will therefore revisit infant f_0 from the perspective of modal and individual non-modal voice qualities (creaky, harsh, breathy, loft, whispery voice) in an effort to more fully understand the f_0 characteristics of early infant vocalisations.

2. Methodology

2.1. Recording and segmentation procedure

A Sony DCR-TRV16E digital video recorder with integrated microphone was used to film four infants (3 female, 1 male) interacting with their caregivers or engaged in solitary play over the first six months of life. The infants were recorded at a sampling rate of 48kHz and 16 bit encoding. Due to the young nature of the subjects (up to 26 weeks) no elicitation of vocalisations was attempted; instead all vocalisations spontaneously produced by the infants during a recording session were later coded, unless background noise was present or the infants had occluded vocal cavities. Each vocalisation was broadly transcribed using a simplified IPA script in the phonetic database software EMU. Each participant's vocalisations were also labelled for voice quality according to auditory-perceptual analysis, supplemented by wide-band spectrograms and time waveforms. The qualities considered for analysis were: harsh voice, creaky voice, whispery voice, modal voice, breathy voice, loft whisper and voiceless. Approximately 10% (1140 vocalisations) of the total corpus of two participants were labelled by an independent rater. Inter-rater reliability for this labelling was calculated at a Cohen's Kappa of 0.76 for phonetic segmentation and 0.80 for voice quality.

2.2. Peculiarities of infant data

Infant vocalisations have characteristics that are quite different from those of adult vocalisations in that they have a wider f_0 range, abrupt f_0 transitions and unique energy distribution patterns over frequencies [6]. Software designed to estimate f_0 routinely experience problems determining the f_0 contour within an infant vocalisation. It will mistakenly determine the

f_0 as either double or half what is correct. [6, 205] says that these types of errors are "often considered to be one of the most significant problems of f_0 estimation." Although this can happen for adult data, it occurs more commonly in the data of infants because of the wider f_0 range.

The extensive use of different voice quality modalities also interferes with the f_0 estimation. Segments displaying creaky or harsh voice have voicing discontinuity because of their production. Although the voicing threshold can be lowered in acoustic analysis software programs such as PRAAT to account for these types of segments, excessive use of this setting can adversely affect the reliability of the f_0 tracking by picking up on 'voicing' that is not actually there. These issues make f_0 pattern estimation of infant vocalisations difficult. Because of these factors, it is important to have a robust methodology to deal with difficulties of working with infant vocalisations. For this reason using the software PRAAT, each spectrogram was individually inspected and the f_0 contour corrected when necessary. The f_0 value extracted by the tracker was then compared to the first harmonic of a Fast Fourier Transform (FFT) for verification. This process provided a robust technique for working with a corpus that included so many aperiodic vocalisations.

Rather than examining vocalisations at the level of the syllable, as used in a number of other studies such as [3] this study calculated the f_0 for each voiced segment at the temporal midpoint. This was done to enable the individual influences of the different voice quality modalities to be examined and analysed. Standard deviation was also calculated in a similar manner. A total of 7,517 segments had their f_0 calculated

3. Results

3.1. Longitudinal mean and standard deviation f_0 trends

Figure 1 presents the combined mean and standard deviation for f_0 across the length of the study. In contrast to [4], no linear developmental trend was evident in the infants' f_0 data. This may be accounted for by the comparatively short length of this present study as [4, 1640] suggest that "two or more years of observation would be necessary to obtain a significant tendency for the f_0 decrease." The reported f_0 decrease per 12 months is so small (between 1.9% and 6.1% in their study) that they would be difficult to detect as a tendency. The present data falls more in line with [1], [2] and [3] who found no changes in mean f_0. In this study the mean f_0 decreased until month 3 and after this point it increased again, see Figure 1. Overall mean f_0 for the 6 month study was 367Hz and is similar to those reported previously [2] and [6]. When looked at individually the four children demonstrated considerable variation, ranging from a mean f_0 of 333Hz to 427Hz.

The standard deviation (SD) also shows a similar pattern decreasing in the initial half of the study before increasing again. The SD also varied across children, ranging from 100Hz to 233Hz. Overall the values were generally higher than those previously reported in the literature [3, 918]. This increase in SD may be attributed to the inclusion of a more diverse corpus of infant vocalisations that includes all sounds the infant produced, especially those with non-modal voice quality.

3.2. Longitudinal voice quality trends

Figure 2 presents the mean f_0 for each voice quality plotted by month. Loft has an extremely high f_0 across the entirety of the study. It has a curve that shows quite a steady decrease in mean f_0 from month 1–4 and then a much greater rate of increase from month 4–6.

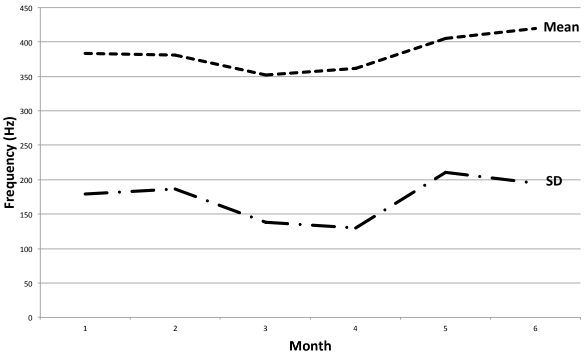


Figure 1: *Combined mean and standard deviation for f_0.*

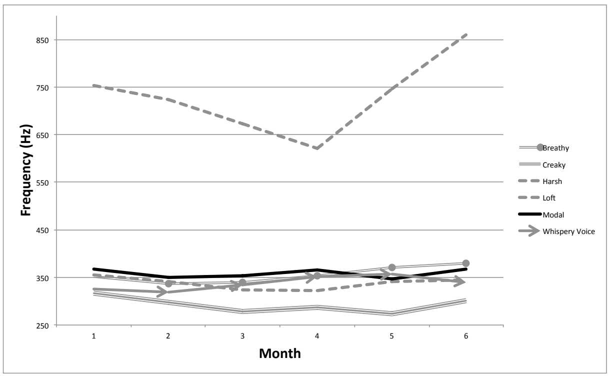


Figure 2: *Longitudinal voice quality mean f_0.*

Figure 3 gives a closer look at the other voice quality modalities. All voice qualities initially experience a decrease in mean f_0 during the first month of the study. For breathy voice this is followed by consistent increases in mean f_0 for the remainder of the study. Whispery voice also experiences increases in mean f_0 for most of the latter part of the study. However, its results must be taken with caution due to the small number of tokens available for analysis (see Table 6). Harsh voice and creaky voice follow a similar pattern to that of loft, though not in the same scale. Decreases in mean f_0 during the first half of the study were again followed by increases in the latter half. Apart from loft, modal voice had the highest mean f_0 for most of the study.

3.3. Variability of voice quality

The coefficient of variation (COV) for f_0 (SD/mean) provides a correction for inter-relatedness, separating variability from absolute values of f_0.[1] When used to examine the variability of voice quality it showed that loft voice had the highest rate of variability, whilst modal had the least. [3] reported that high mean f_0 tends to correspond to high variability in absolute values of f_0. This proves to be the case with loft voice exhibiting the highest amount of variability as well as the highest mean f_0. However modal voice always exhibited the lowest amount of variability, despite the fact it maintained the second high-

[1] The number of tokens can affect the COV. Whispery voice was a small proportion of the data set (see Table 4) therefore its high level of variability should not be given too much importance.

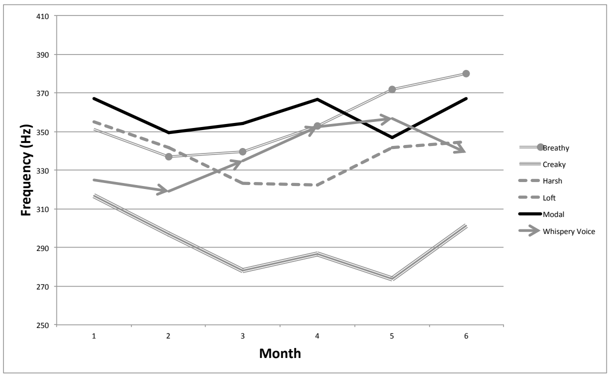


Figure 3: *Longitudinal voice quality mean f_0 (excluding loft).*

est mean f_0. When the data were examined by month, modal voice had the least amount of variability across the entirety of the study. This shows that infants are potentially regulating the amount of f_0 variability that they are producing in different voice qualities. It is important to note that the dominant voice quality in English (modal voice) has a demonstrably higher degree of control being exercised over the use of f_0 and that this is apparent even from the first month of life.

Table 1: *Mean COV for different voice qualities.*

Breathy	Creaky	Harsh	Loft	Modal	Whispery Voice
0.30	0.35	0.42	0.44	0.23	0.40

3.4. Statistical Analysis

A linear mixed effects model was performed which incorporated both random and fixed effects. This analysis was particularly appropriate for spontaneous infant 'speech' because of its flexibility in handling missing values and unmatched numbers of tokens in the individual participants. In addition, mixed effects models offer the advantage of providing insights into the full structure of the data by examining fixed and random effects simultaneously [7]. Analyses were carried out using the R statistical computing software [8]. In the model, the dependent variable was f_0 (transformed into bark for normality). The independent variables in the model included one random-effect factor (subject) and three fixed effect factors (phonetic category, perceptual voice quality and month). Only the results for voice quality and month will be discussed in this paper. A model where there was interaction between perceptual voice quality and month performed significantly better ($\mathcal{X}^2 = 81.094, df = 30, p < 0.001$) than one without interaction. A Tukey post-hoc comparison accounting for interaction was then conducted in order to ascertain significance.

There were some significant effects evident between individual months. Table 2 shows these interactions. These interactions show that there is a statistically significant difference between the lowest f_0 values (months 2, 3 and 4) and the highest f_0 values (months 5 and 6). However across the entirety of the study there are no significant changes in f_0. As such no longitudinal trend of increasing or decreasing f_0 can be ascertained. When considering voice quality there are some significant effects. The f_0 of loft segments and creaky voice segments are significantly ($p < 0.001$) different from all other voice qualities. Creaky voice has a lower f_0 than all other voice qualities, whilst loft has higher f_0 than every other voice quality. Harsh voice also has a significantly different f_0 than a number of the other voice qualities. These interactions are shown in Table 3.

Table 2: *Fundamental frequency interaction by month.*

		Month					
		1	2	3	4	5	6
Month	1	▨					
	2		▨			*	***
	3			▨		**	***
	4				▨	*	***
	5		*	**	*	▨	
	6		***	***	***		▨

Statistical significance
($* = p < 0.05$), ($** = p < 0.01$), ($* * * = p < 0.001$)

Table 3: *Fundamental frequency interaction by voice quality.*

	Breathy	Creaky	Harsh	Loft	Modal	Whispery Voice
Breathy	▨	***	***	***		
Creaky	***	▨	***	***	***	***
Harsh	***	***	▨	***	***	
Loft	***	***	***	▨	***	***
Modal		***	***	***	▨	
Whispery Voice		***		***		▨

Statistical significance
($* = p < 0.05$), ($** = p < 0.01$), ($* * * = p < 0.001$)

4. Discussion

The results in this study serve to provide further clarification regarding the f_0 trends in typically developing infants. The mean f_0 for the infants of 367Hz was similar in value to a number of previous studies including: [9] and [10]. It is almost identical to that reported by [6]. However it is quite different from that reported by [11] and [12]. The mean f_0 results from these studies were quite high (529Hz and 450Hz respectively). A reason for the lower values found in the present study is the inclusion of all infant vocalisations produced during the recording sessions. This included a large proportion of sounds (see Table 4) produced with extremely low f_0 such as creaky voice segments.

In terms of developmental trends, this present study observed mean f_0 fluctuations month to month and differing patterns between infants. Such individual variation was also observed by [4] and [9]. However longer range trends as seen in [12], [13], and [14] were not in evidence in this data. Instead the results of this study parallel those of the [15] study where no consistent increase or decrease was observed for mean f_0 between 0 months and 6–9 months. The results are also similar to [11], which reported that the mean f_0 decreased between zero and one month and then increased and became stable at 2–4 months. These latter two studies also utilised a methodological approach in which non-modal vocalisations were accepted for analysis and this may have played a role in the comparability of results. The lack of an overall decreasing trend is notable due to the anatomic changes occurring during the timeframe of this study. A decrease in f_0 would be hypothesised due to the lengthening of the vocal tract in both the oral and laryngeal dimensions [16]. However [14] did not find a decrease until after a period of relative stability during the first year. Although a longer and larger scale study would be needed to further clarify the overall longitudinal trends of f_0, this present study does help to reveal f_0 changes over a short time period.

The most significant finding of this current study is the impact of voice quality on infant f_0. Loft, creaky voice and harsh voice are significantly different from other voice qualities based on f_0 alone. It is suggested that the developmental pattern evidenced here for loft vocalisations, played a role in determining the overall contour of the data as seen in Figure 1. While loft vocalisations as a whole only make up a small proportion of each months' productions (see Table 6), the high variability in mean loft f_0 have influenced the overall developmental contour. Harsh and creaky voice also displayed similar developmental patterns. Although their variability was not as great as lofts, they comprised a larger proportion of the data set. Together these three voice qualities (loft, creaky voice and harsh voice) have the largest mean coefficient of variation.

Although infants still have variable f_0 (seen in this study as high mean and SD measures) in the first six months of life, in terms of f_0 variability (SD/mean), modal voice demonstrated the highest degree of control. This control occurs within the first month of life and remains for the entirety of the study. Even whilst significant changes are occurring in the anatomic-physiological structure of the infants' vocal tract and respiratory system, an infant is able to regulate the degree of f_0 variability so as to best mimic the dominant surrounding voice quality. This suggests increasing control of the larynx and vocal fold responsible for voicing.

It also suggests that previous studies' comparability issues due to methodological differences continue to need to be addressed. By focusing on just one aspect of one of the areas that [3] identified, it has been shown how voice quality contributes to the f_0 characteristics of infant vocalisations.

Table 4: *Monthly proportion of vocalisations produced with each auditory-perceptual voice quality*

Voice	**Month**					
Quality	1	2	3	4	5	6
Breathy	7.9%	10.5%	15.9%	10.2%	6.6%	10.2%
Creaky	21.3%	13.2%	11.8%	12.8%	14.2%	14.0%
Harsh	23.3%	15.8%	17.5%	24.7%	21.4%	24.9%
Loft	6.0%	6.3%	3.9%	3.7%	7.3%	4.6%
Modal	6.2%	9.4%	17.4%	25.8%	24.6%	22.4%
Voiceless	29.1%	39.2%	32.2%	20.0%	22.4%	20.7%
Whisper	2.1%	2.2%	0.4%	0.9%	0.9%	0.8%
Whispery Voice	4.1%	3.4%	0.9%	1.9%	2.5%	2.4%

5. Conclusions

This paper provides additional insight into the developmental trends evident in pre-babbling infants' vocalisations. Across the course of the study, no overall decrease or increase was evident in mean f_0. In addition, the overall mean and SD were similar to those found in previous studies. By including vocalisations produced with non-modal phonation, their impact on the f_0 data was able to be discerned. Individual voice qualities were able to be distinguished from one another on the basis of f_0. Different voice qualities also displayed varying amounts of variability (SD/f_0) across the entirety of the study, with modal voice showing the least variation. Whilst the growth of the vocal tract seems to have limited amounts of impact on the mean values of f_0, it does have a role in displaying the increasing control infants have over the processes used for voicing. Infants are able to regulate the degree of f_0 variability even whilst anatomic changes are occurring. The results presented in this paper support the notion that the f_0 of infant vocalisations provide insight into how an infant learns to exercise vocal control and that voice quality is a useful category through which to investigate these developments.

6. References

[1] Delack, J. B., & Fowlow, P. J. (1978). The ontogenesis of differential vocalization: development of prosodic contrastivity during the first year of life. In The development of communication (pp. 93–110). New York: John Wiley & Sons.

[2] Whalen, D. H., Levitt, A. G., Hsiao, P. L., & Smorodinsky, I. (1995). Intrinsic F0 of vowels in the babbling of 6-, 9-, and 12-month-old French-and English-learning infants. Journal of the Acoustical Society of America, 97, 2533–2539.

[3] Iyer, S. N., & Oller, D. K. (2008). Fundamental frequency development in typically developing infants and infants with severe-to-profound hearing loss. Clinical Linguistics & Phonetics, 22(12), 917–936.

[4] Amano, S., Nakatani, T., & Kondo, T. (2006). Fundamental frequency of infants and parents utterances in longitudinal recordings. The Journal of the Acoustical Society of America, 119(3), 1636.

[5] Buder, E. H., Chorna, L. B., Oller, D. K., & Robinson, R. B. (2008). Vibratory Regime Classification of Infant Phonation. Journal of Voice, 22(5), 553–564.

[6] Nakatani, T., Amano, S., Irino, T., Ishizuka, K., & Kondo, T. (2008). A method for fundamental frequency estimation and voicing decision: Application to infant utterances recorded in real acoustical environments. Speech Communication, 50(3), 203–214.

[7] Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. Journal of Memory and Language, 59(4), 390–412.

[8] Team, R. C. (2012). R: A language and environment for statistical computing. Vienna: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org

[9] Laufer, M. Z., & Horii, Y. (1977). Fundamental frequency characteristics of infant non-distress vocalization during the first twenty-four weeks. Journal of Child Language, 4(02), 171–184.

[10] Robb, M. P., & Saxman, J. H. (1989). Vocal fundamental frequency characteristics during the first two years of life. Journal of the Acoustical Society of America, 85(4), 1708–1717.

[11] Sheppard, W. C., & Lane, H. L. (1968). Development of the Prosodic Features of Infant Vocalizing. Journal of Speech and Hearing Research, 11(1), 94.

[12] Kent, R. D., & Murray, A. D. (1982). Acoustic features of infant vocalic utterances at 3, 6, and 9 months. Journal of the Acoustical Society of America, 72(2), 353–365.

[13] Bennett, S., & Bennett, S. (1983). A 3-Year Longitudinal Study of School-Aged Children's Fundamental Frequencies. Journal of Speech and Hearing Research, 26(1), 137.

[14] Vorperian, H. K., & Kent, R. D. (2007). Vowel Acoustic Space Development in Children: A Synthesis of Acoustic and Anatomic Data. Journal of Speech, Language and Hearing Research, 50(6), 1510–1545.

[15] Prescott, R. (1975). Infant cry sound; developmental features. Journal of the Acoustical Society of America, 57, 1186–1191.

[16] Vorperian, H. K., Kent, R. D., Lindstrom, M. J., Kalina, C. M., Gentry, L. R., & Yandell, B. S. (2005). Development of vocal tract length during early childhood: A magnetic resonance imaging study. Journal of the Acoustical Society of America, 117(1), 338–350.

3-year-olds Produce Pitch Contours Consistent with Mandarin Tone 3 Sandhi

Nan Xu Rattanasone[1], *Ping Tang*[1], *Ivan Yuen*[1], *Liqun Gao*[2], *& Katherine Demuth*[1]

[1]Department of Linguistics, Center for Language Sciences, Macquarie University, Australia
[2]Beijing Language and Culture University, China

nan.xu@mq.edu.au; ping.tang1@students.mq.edu.au; ivan.yuen@mq.edu.au; gaolq@blcu.edu.cn; Katherine.demuth@mq.edu.au

Abstract

Few studies have examined the acoustic properties of children's tonal productions. This study examined the productions of lexical tone and Tone 3 sandhi from 27 Mandarin-speaking 3-year-olds. Mandarin is a language with four lexical tones, T1 high level, T2 rising, T3 dipping and T4 falling, and both Full and Half Tone 3 sandhi processes. The results showed that 3-year-olds are producing all four lexical tones and tone sandhi, at least for lexicalized forms.

Index Terms: Lexical tone acquisition, tone sandhi, Mandarin

1. Introduction

Languages that have lexical tone make use of manipulations in pitch height and pitch contours to change the meanings of words. Whereas in English rising and falling pitch contours over a single word are typically associated with prosodic information such as focus, in lexical tone languages these changes can alter the meaning of the word. Mandarin is one of the best studied lexical tone languages, with the largest number of speakers. However, despite the pervasiveness of tone languages, relatively little is known about the acquisition of tones compared to other phonological contrasts (i.e., vowels and consonants). Even less is known about the acquisition of tone sandhi, where the surface tone changes depending on the tone context, e.g., Mandarin Tone 3 sandhi.

Mandarin has a four-tone system with one level and three contour tones; Tone 1 (T1) is High Level (‘mā’: *mother*), Tone 2 (T2) is Rising (‘má’: *hemp*), Tone 3 (T3) is Dipping (‘mǎ’: *horse*), and Tone 4 (T4) is Falling (‘mà’: *reprimand*). While all four tones appear in the productions of Mandarin-speaking children by the 1-word stage of development, confusion between Tones 2 and 3 (Rising and Dipping tones) continues into the 2/3-word stage of development, finally disappearing as longer sentences are produced[1]. Compared to other tone languages, acquisition of the Mandarin tonal system appears to be more protracted[2]. For example, for Cantonese, a language with a six-tone system, children have acquired all tones by the age of 2[3]. This includes tones with very similar pitch contours, e.g., three level tones (high, mid & low) and two rising tones (high & mid). The same early acquisition of the full tone inventory is also observed in Thai, a language with 5 tones, where all tones were present by the 2-word stage[4]. Thai also has three level tones (High, Mid, & Low). Thus, the similarities between pitch contours and larger tone inventory does not appear to delay tone acquisition.

The delayed mastery of the tonal system in Mandarin might be related to tone change processes which do not occur in Cantonese or Thai. Mandarin Tone 3 sandhi is a tone change process that occurs when the underlying tone changes to different surface forms in various tone contexts. When two T3 occur in succession (T3-T3), the first becomes a rising tone and is called the *full sandhi*, but when T3 is followed by any other tone, it falls in pitch; this is called the *half sandhi*. Previous studies have reported that Tone 3 sandhi begins to be acquired by the 2/3-word stage of development, with few errors[1]. It is after this point that the confusion between T2 and T3 is resolved. One reason for this might be that in the *full sandhi* form T3 has a rising contour like T2, and children need to first learn the sandhi rule before understanding the difference between lexical T2 and Tone 3 sandhi. In Bantu languages such as Sesotho, where lexical and grammatical tone interact, tone sandhi processes are acquired only by 3 years or later, as children learn more about the grammar of the language[5].

Acquisition studies typically report only briefly on Tone 3 sandhi, with results typically based on auditory impressions of the children's productions. One study reported that tone sandhi emerges early, at around 2 years[6], but studies have not examined the acoustic realization of children's tone sandhi productions.

To examine this issue, productions of lexical tone and contexts where tone sandhi should occur were collected from Mandarin monolingual 3-year-olds using an elicitation task. Since so little is known about the acoustic realization of children's use of tone sandhi, real words were used, testing lexicalized forms of Tone 3 sandhi.

2. Methods

2.1. Participants and design

A total of 27 3-year-olds (13 boys; 14 girls) participated in the study. The mean age of the children was 3;10 (range 3;6 – 3;11). They were recruited in Beijing from the preschool associated with the Beijing Language and Culture University. The study was conducted in accordance with the ethics protocol approved by Macquarie University's Human Ethics Panel. All child participants received fun stickers for their participation and the preschool received book donations for all children to use at the center.

A within-subjects design was used, and all children were tested on all target items.

2.2. Stimuli

A total of 34 high-frequency disyllabic words were used (Table 1). To elicit the lexical tones, 12 disyllabic words with T1, T2, and T4 as the first syllable and T1 to T4 as the second syllable were chosen. It was not possible to find enough words all beginning with T1 to avoid tone co-articulation effects, so a variety of tone contexts were used. It was also not possible to avoid words ending in nasal /n/ and /ŋ/ codas, which can have the effect of lowering the pitch of the syllable.

To elicit Full sandhi, five disyllabic T3-T3 words were chosen. For Half sandhi five words, each with T3 as the first syllable and T1, T2, and T4 as the second syllable were chosen, yielding 20 stimulus words for tone sandhi contexts. An additional two practice items in the form of T3-T3 (a puppy and a pony) were used at the beginning of each session but were not analyzed.

Most syllables were CV in form, and where possible contained a stop or fricative/affricate onset to facilitate acoustic coding. However, a few contained a lateral or nasal onset, and some contained a nasal coda. Two versions of the test were created, each with a different randomization for the order of presentation of the words.

Table 1. *List of Disyllabic Stimuli Words*

	Tones	Pinyin	Meaning
Practice	T3T3	xiao-gou	puppy
	T3T3	xiao-ma	pony
Full sandhi	T3T3	lao-shu	mice
	T3T3	lao-hu	tiger
	T3T3	xiao-niao	chick (bird)
	T3T3	yu-san	umbrella
Half sandhi	T3T1	xiao-mao	kitten
	T3T1	jian-dao	scissors
	T3T1	kao-ya	Peking duck
	T3T1	yu-yi	raincoat
	T3T1	bing-gan	biscuit
	T3T2	er-huan	earing
	T3T2	kou-hong	lipstick
	T3T2	cai-hong	rainbow
	T3T2	cao-mei	strawberry
	T3T2	xiao-niu	calf
	T3T4	li-wu	present
	T3T4	shou-tao	gloves
	T3T4	tu-dou	potatoes
	T3T4	kong-que	peacock
	T3T4	tan-ke	tanker
Lexical Tone	T1T1	xi-gua	watermelon
	T1T2	ying-tao	cherries
	T1T3	ban-ma	zebra
	T1T4	ji-dan	egg
	T2T1	long-xia	lobster
	T2T2	liang-xie	sandals
	T2T3	ping-guo	apple
	T2T4	qin-cai	celery
	T4T1	li-zhi	lychee
	T4T2	qi-qiu	balloon
	T4T3	chi-bang	wings
	T4T4	da-xiang	elephant

2.3. Materials

A total of 34 non-proprietary photographic images representing each of the 32 test and 2 practice items were selected from google images. The images were presented one at a time using Microsoft PowerPoint 2013 delivered on an Apple iPad 2. The recordings were collected using a Zoom H2 digital voice recorder with lapel mic and the recordings were exported as PCM files.

2.4. Procedure

Testing was conducted in a quiet area in the preschool. Each child was greeted by the native Mandarin-speaking experimenter. The task was explained as a picture naming game where children named the pictures on the screen and receive stickers for playing the game. Two practice trials were given and for children who could not provide an answer after three prompts, the experimenter provided the answer, e.g., “puppy”. The child was then asked to repeat the label before moving to the next item. The children were encouraged to provide answers independently during the practice trials. All children were able to perform the elicitation task, however, there were two items where a majority of the children could not identify, i.e., T3-T2 *rainbow* and T2-T1 *lobster*. For these items, the experimenter produced the items but the imitations from the children were not analyzed.

2.5. Data Analysis

The productions were acoustically coded in Praat[7] by a trained coder who is a native speaker of Mandarin. The tones were extracted from the vocalic portion of the syllable (and nasal if present). The vocalic portion was identified from the onset of higher formants to the cessation of higher formants in the first syllable and offset of voicing as indicated by the onset of the second syllable. In cases where the second syllable had a nasal onset, anti-resonance and simplification of the waveform was used to identify the onset of the second syllable. Using Praat[7], the average F0 for each syllable were extracted in 10 equal steps.

3. Results

To determine that 3-year-olds are producing lexical tones, analysis of the F0 changes over time were conducted. If children are specifying the four tones, there should be significant linear (rising or falling) or quadratic (dipping) trends emerging over time for all four tones. For Tone 3 sandhi, F0 changes were analyzed separately for Full and Half sandhi contexts. Analysis was conducted over the first syllable only where tone sandhi was expected to manifest. If children are producing Tone 3 sandhi then F0 changes should be different across these sandhi contexts with linear trends (rising vs. falling) for Full and Half sandhi.

3.1. Lexical Tone

A linear mixed effects regression model (LMEM) was conducted with F0 as the dependent variable. Polynomial equations (up to quadratic) were fitted for the 10 Time points with Tone type (T1 to T4) as the other fixed factor. Random intercepts for each child and item were fitted for the effect of tone (F0 ~ poly(Time, 2)*Tone + (1 | Child) + (1 | Word)). See results on Table 2 and Figure 1.

The results show a main effect of Tone type. Mean F0 for T2 (t = -5.991, p < .01) and T3 (t = -13.164, p < .01) are significantly lower than T1 (the referent category). The interactions show a significant linear trend for T2 (t = 2.832, p = .005) and a positive effect on the intercept, suggesting a linear increase over time (rising contour). There were significant linear (t = -5.717, p < .01) and quadratic (t = 7.013, p < .01) trends for T3. The negative effect on the intercept for the linear trend and positive effect for the quadratic trend suggest that T3 decreased before rising over time (dipping contour). The significant linear trend for T4 (t = -7.274, p < .01), and the negative effect on the intercept suggest a linear decrease over time (falling contour). These F0 changes over time (rising for T2, dipping for T3 and falling for T4) are consistent with the expected tone contours for each tone.

Table 2: *Results for F0 of lexical tone across 10 time points* (Kenward-Roger approximations were made for degrees of freedom; ***p* < .01)

	Estimate	Std. Error	*t* value	*p* (KR)
(Intercept)	286.736	9.211	31.130	0.000
Main Effects				
poly1	24.047	146.413	0.164	0.870
poly2	63.949	122.340	0.523	0.601
T2	-43.331	7.232	-5.991	0.000 **
T3	-95.168	7.229	-13.164	0.000 **
T4	-0.774	7.229	-0.107	0.915
Interactions				
poly1:T2	448.293	158.306	2.832	0.005 **
poly2:T2	299.838	158.280	1.894	0.058
poly1:T3	-902.687	157.886	-5.717	0.000 **
poly2:T3	1107.326	157.886	7.013	0.000 **
poly1:T4	-1148.486	157.886	-7.274	0.000 **
poly2:T4	115.326	157.886	0.730	0.465

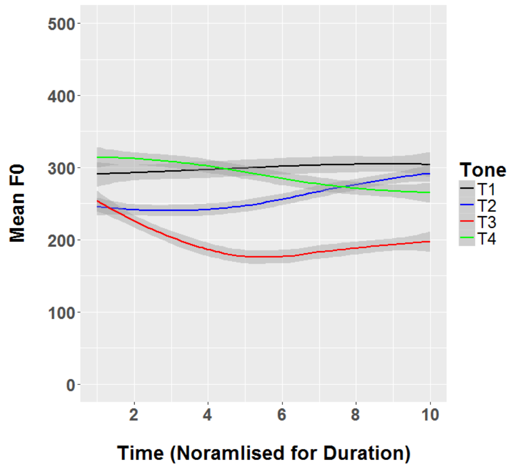


Figure 1. *Mean F0 at 10 time points for the four lexical tones*

3.2. Tone 3 Sandhi

To examine whether 3-year-olds are distinguishing between Full and Half sandhi, two LMEMs estimating polynomial trends over time were conducted separately for Full and Half sandhi contexts. The same model and factors were used as above after removing Lexical Tone as a factor. See results on Table 3 and Figure 2.

The results show significant linear and quadratic trends for both Full and Half sandhi contexts. For Full sandhi, the positive effect on the intercept for both the linear (t = 7.227, p < .01) and quadratic (t = 3.803, p = .001) trends suggest a steep rise, then slowing over time. For Half sandhi, the negative effect on the intercept for the linear trend (t = -12.729, p < .01) and positive effect for the quadratic trend (t = 5.913, p < .01) suggest a falling then slight rising (dipping) contour over time. While the rising contour is consistent with Full sandhi, the dipping contour is not consistent with Half sandhi which should have a falling contour.

Table 3. *Results for F0 of full and half sandhi contexts across 10 time points* (**p < .01)

	Estimate	Std. Error	*t* value	*p* (KR)
Full Sandhi				
(Intercept)	282.895	10.316	27.423	0.000
poly1	529.127	73.217	7.227	0.000**
poly2	160.272	42.148	3.803	0.001**
Half Sandhi				
(Intercept)	236.665	6.221	38.041	0.000
poly1	-627.542	49.301	-12.729	0.000**
poly2	244.237	41.305	5.913	0.000**

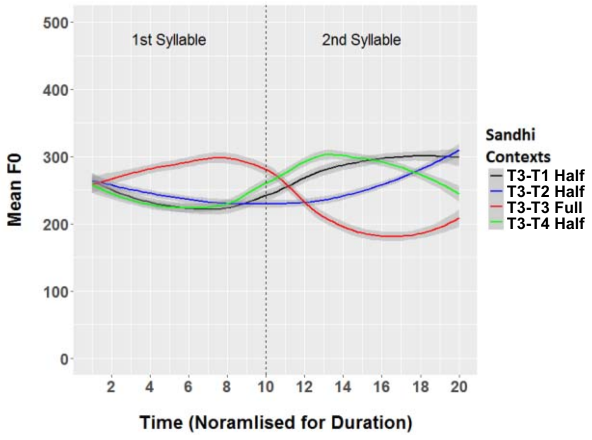


Figure 2. *Mean F0 at 20 time points for disyllabic word contexts eliciting full and half sandhi*

4. Discussion

The aim of this study was to examine the acoustic realizations for lexical tones and Tone 3 sandhi contexts in the productions of 3-year-olds. First, the analysis of lexical tone production suggested that children are producing contours consistent with that of the four lexical tones, i.e., level for T1, rising for T2, dipping for T3 and falling for T4. Secondly, the analysis of Tone 3 sandhi contexts suggested that they are also producing sandhi forms in these high frequency known words: the *full sandhi* had a rising contour while the *half sandhi* had a dipping contour. The *full sandhi* is consistent with the expected contour, however the results for *half sandhi* suggested that 3-year-olds might be producing the underlying lexical T3. Overall, the results suggest that, in addition to having a well specified lexical tone space, 3-year-olds are also producing *Full sandhi*.

These results raise several questions for future research, such as how adult-like the children's productions are, not only in terms of the contours but also on other measures. For example, T3 is typically associated with creaky voice in adult productions but it is unclear if children also use this cue. Adults also use a variety of acoustic cues to signal lexical tones, including differences in pitch onset, offset, range and turning points. Knowing whether children are using similar cues can better inform our understanding of what children are encoding when learning the tone space. There is also considerable individual variation (Figure 3), showing that some children have good tonal representations and tone sandhi productions (child 11 & 12) while others show poor tone separation and sandhi production (child 3, 16 & 22).

5. Conclusions

Mandarin-speaking 3-year-olds can produce Tone 3 sandhi in *full sandhi* contexts on familiar/known lexicalized items. Future studies should examine children's ability to apply tone change processes to novel words.

6. Acknowledgements

This work is funded by the Macquarie University Research and Development Grant [1547620] awarded to the first author and the ARC Laureate Fellowship [FL130100014] awarded to the last author. We thank Dr. Ivan Yuen, members of the Child Language Lab, and the two anonymous reviewers for their helpful comments.

7. References

[1] Li, C. N., & Thompson, S. A., "The acquisition of tone in Mandarin-speaking children", Journal of Child Language, 4(02): 185–199, 1977.

[2] Wong, P., "Acoustic characteristics of three-year-olds' correct and incorrect monosyllabic Mandarin lexical tone productions", Journal of Phonetics, 40(1), 141–151, 2012.

[3] So, L. K. H., & Dodd, B. J., "The acquisition of phonology by Cantonese-speaking children", Journal of Child Language, 22: 473–495, 1995.

[4] Tuaycharon, P., "The phonetic and phonological development of a Thai baby: From early communicative interaction to speech", Unpublished PhD Thesis, University of London, 1977.

[5] Demuth, K., "Issues in the acquisition of the Sesotho tonal system". Journal of Child Language, 20: 275-301, 1993.

[6] Hua, Z., & Dodd, B., "The phonological acquisition of Putonghua (Modern Standard Chinese)", Journal of Child Language, 27: 3–42, 2000.

[7] Boersma, P., & Weenink, D., "Praat: doing phonetics by computer [Computer program]", Version 6.0.18, retrieved 23 May 2016 from http://www.praat.org/.

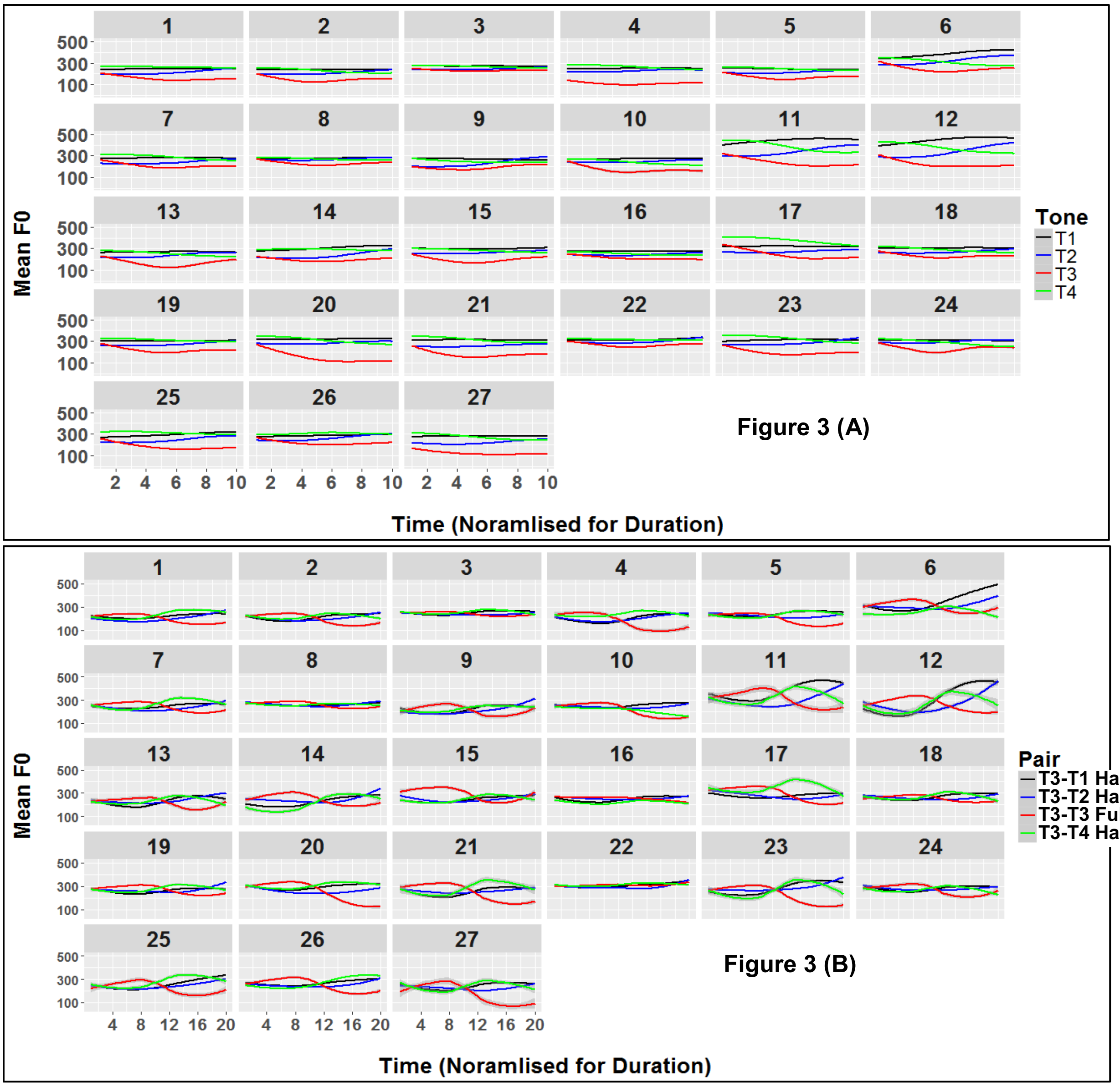


Figure 3. (A) *Mean F0 at 10 time points of lexical tones for each child*, (B) *Mean F0 at 20 time points for disyllabic words for sandhi contexts for each child*

The role of positive affect in the acquisition of word-object associations

Nicole M. Traynor[1,2,3], *Karen E. Mulak*[1,2], *Rachel Robbins*[1,3], *Gabrielle Weidemann*[1,2,3], *Paola Escudero*[1,2]

[1] MARCS Institute for Brain, Behaviour and Development, Western Sydney University, Australia
[2] Centre of Excellence for the Dynamics of Language, Australian Research Council, Australia
[3] School of Social Sciences and Psychology, Western Sydney University, Australia

n.traynor@westernsydney.edu.au

Abstract

Learning to associate words to their meaning is a difficult task. Early word learning may be aided by the way in which adults talk to infants. Infants prefer infant-directed speech (IDS) over adult-directed speech (ADS) [1]–[3], and evidence suggests the positive affect inherent to IDS drives this preference [4]. Infants can form word-object associations in IDS [5], [6], but we do not know what role affect plays on word learning. We tested 19-month-olds' learning of word-object pairings when words were taught in a positive or neutral affect in ADS. No evidence of word learning was found. Results and future research implications are discussed.

Index Terms: word learning, infancy, eyetracking

1. Introduction

It appears that early word learning is aided by the way in which adults talk to infants. When parents and caregivers speak to infants they spontaneously modify the way they talk. This form of speech is referred to as infant directed speech (IDS). A number of studies have shown that infants demonstrate a preference for IDS over adult directed speech (ADS;[1]–[3]); however, what drives this preference is less understood. Characteristics of IDS include shorter sentences, longer pauses, greater pitch variation, slower tempo, a higher fundamental frequency, and increased repetitions [7]–[9] Trainor, Austin, and Desjardins [10] argue that emotional expression is the primary determinant of IDS, suggesting that as infants cannot understand the words, emotional prosody is an important aspect of what is communicated. They suggest that reported differences between IDS and ADS are in part the result of comparing more positive IDS with less positive ADS.

Singh, Morgan, and Best [4] attempted to disentangle infants' preferences for IDS, ADS, and affect over a number of experiments and found that when affect was held constant, 6-month-old infants did not show a preference for IDS over ADS. Importantly, infants demonstrated a preference for ADS when it was presented with more positive affect than the IDS. These results provide strong evidence that it is the positive affect inherent within IDS that underlies infants' preference.

Not only do infants prefer IDS but it also seems to help them learn. IDS facilitates infants' speech perception performance at 6–8 months and 10–12 months [11], word segmentation at 6.5–8.5 months [12], word recognition at 7.5 months [13], and phoneme categorisation at 12 months [14]. More recently, it has been demonstrated that infants at 17 months [5] and 21 months [6] can make accurate word-object associations when familiarised in IDS; however, they cannot do the same when familiarised in ADS.

Infants' preference for IDS has also been demonstrated to extend to individuals who have been previously speaking in IDS. In their study, Schachner and Hannon [15] presented 5-month-old infants with two videos, one in which a woman spoke using IDS, and another with a different woman speaking with ADS. A visual preference test was then presented to the infants, with a static photo of one of the familiar faces the infant had just viewed, presented side by side on the screen with the photo of a novel woman's face. When the familiar face presented was the one that used IDS, infants preferred to look at the familiar face over the novel face. When the familiar face presented was the one that had used ADS, a preference was shown for the novel face. These results suggest that the pairing of IDS and an individual can lead to a change in the preference for the individual by infants. Therefore, if it is the positive affect within IDS that draws infants' attention, affect may also aid in learning and/or forming other associations.

Affect has also been shown to aid memory and attention in infants, and specifically to promote language acquisition. Infants at 7.5-months can recognise words in positive and neutral affect only when affect is matched across familiarisation and testing, while infants at 10.5-months are able to recognise words across variations in affect [16]. When a word was familiarised in varied affect (happy, sad, neutral, angry, and fearful), 7.5-month-old infants' are capable of recognising the word in fluent speech [17].

The aim of the current study was to determine if positive affect can aid infants' learning of word-object associations. To answer this question, infants were familiarised to four novel objects and their associated novel words, two of which were named using positive affect, and two named with neutral affect. As previous studies have demonstrated that infants at 17- [5] and 21- [6] months can learn word-object associations in IDS but not in ADS, 19-month-olds were used in this study. Ability to form associations between the words and objects was assessed using preferential looking at the named object. Following this, infants' preferences for the novel objects was measured using an object-reaching task. Due to infants' preference for positive affect, we hypothesised that infants would attend more to the objects named with positive affect, than those named with neutral affect. Therefore, it was also hypothesised that infants would form more accurate word-object associations for objects named with positive affect than those named with neutral affect. It was further hypothesised that the pairing of objects with positive affect will result in a change of preference for these objects, and that infants will therefore prefer to play with objects named with positive affect than those named with neutral affect.

2. Method

2.1. Participants

Participants were twenty-one 19-month-olds (*M* = 18.8 months, 11 males), born full-term, from Australian English (AusE)-speaking households in Sydney, Australia. Exposure to non-native languages and non-AusE accents ranged from 0 to no more than 12 hours per week, as indicated by parental report. All parents provided informed consent in accordance with the Western Sydney University Human Research Ethics Committee, and were recruited via pregnancy and baby fairs and magazine advertisements.

2.2. Stimuli and Apparatus

Four novel objects were created from polystyrene and felt. For instructional purposes, two objects familiar to infants were also used: images of a baby and a cat (see *Figure 1*). These are words that occurred in the expressive vocabulary of 75% and 66%, respectively, of 19-month-old infants according to the MacArthur-Bates Communicative Development Inventory (MCDI): Words and Sentences form designed to be used with children aged 16- to 30-months-old [18]. Four novel words were used: a minimal consonant pair, “bon” and “pon”, and a minimal vowel pair, “deet” and “dit”.

Eight familiarisation videos consisted of an AusE-speaking female adult speaking directly to the camera, with positive affect in facial expression and voice (cf., [19]). In another set of eight videos, the same adult woman is seen speaking with neutral affect and neutral expression (see *Figure 2*). Two carrier sentences were used (one per video): “This is the [word]” and “Look at the [word]” (cf., [20]), while the woman held up the associated novel object. The audio for each video was edited to ensure the novel word was the same across all videos within each affect, and the intensity scaled to 65dB.

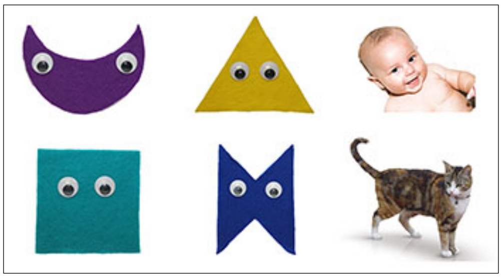


Figure 1: *The four novel objects and two familiar objects*

The videos were rated by undergraduate students for affective valence using a 7-point Likert scale ranging from 1 (*very negative*) to 7 (*very positive*) with neutral emotion represented by the intermediate point. The difference between the positive and neutral videos was found to be highly significant ($p < .001$).

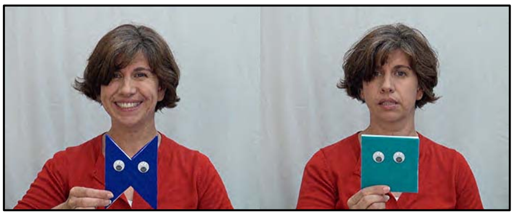


Figure 2: *Screen shots from videos used during familiarisation. Two objects were named using obvious positive affect in facial expression and voice (left), and two objects named using neutral facial expression and voice (right).*

2.3. Procedure

Infants were tested individually in three phases: a familiarisation phase, a test phase, and behavioural choice test phase. During the familiarisation and test phases, infants were seated on their parent’s lap approximately 60-70 cm from the screen, with a Tobii X120 eyetracker positioned below the screen to record infants’ eye gaze on the screen. The parent wore headphones playing words over music to mask the audio from the experiment. Using Tobii Studio, the experimenter, located in an adjacent control room, performed a 9-point calibration. E-Prime 2.0 was used to present the stimuli for the familiarisation and testing phases.

During the familiarisation phase, an “attention-getter” – a short video containing an animated character with a non-linguistic sound – was used before each trial to ensure the infant’s gaze was fixated on the centre of the screen. The experimenter commenced each trial once the infant’s gaze was fixed on the attention-getter. Infants were presented with a total of 24 trials containing eight different videos, presented in three sets in pseudo-randomised order. During each set, videos containing the same object, affect, or carrier sentence were presented no more than twice in a row.

During the testing phase, objects were presented side-by-side on the screen, and the infant was instructed to look at one of the objects with three repetitions of the sentence “Look at the [word]” (see *Figure 3*). The first two trials contained a picture of a cat on one side and a baby on the other, to teach the infants the task and to ensure they could perform the word-object association with a familiar word (cf., [21]). Following this, two sets of 10 trials were presented, with each novel word and one familiar object presented as the target once on each side during each set.

During the behavioural choice phase, a second experimenter who was blind to the conditions entered the testing room and requested that the parent close his or her eyes to ensure that he or she did not guide the infant’s preference. The experimenter then presented two of the objects on a tray to the infant. The auditory word referring to each object formed a minimal pair, in which one object was named previously with positive affect, and the other named with neutral affect. The infant was then asked to choose which object they would like to play with. This procedure was repeated with the second pair of novel objects. Preference was measured based on which object was reached for first, as noted by the experimenter.

3. Results

A paired-samples *t*-test compared percent fixation to the target (i.e., named) image for target objects between test trials in which target objects familiarised in positive affect (*M* = 46.07, *SD* = 13.01), and test trials in which target objects were familiarized in neutral affect (*M* = 47.48, *SD* = 10.83). There was no difference in target fixation between familiarization affect *t*(20) = -.34, *p* = .740.

One-sample *t*-tests compared all combinations of the target and distractor objects: familiar object vs familiar object; familiar object vs positive object; familiar object vs neutral object; positive object vs positive object; positive object vs neutral object; and neutral object vs neutral object. Target fixation was above chance only when both the target and distractor was a familiar object (i.e., the baby and cat), *t*(20) = 3.06, *p* = .006, indicating that although infants did not learn the novel word-object pairings, they could still perform the task. Means and standard deviations for all comparisons are presented in Table 1.

A chi-square test of goodness-of-fit was conducted on the behavioural choice test, and no statistical significant difference was found for preference for objects familiarised in positive affect over neutral affect for the first choice, χ(19) = 0.8, *p* = .371 or the second choice χ(19) = 0, *p* = 1.000.

Comparison	*M*	*SD*
Familiar object vs familiar object	62.21	18.30
Familiar object vs positive object	55.54	17.57
Familiar object vs neutral object	56.06	20.30
Positive object vs positive object	56.97	26.14
Positive object vs neutral object	46.03	7.80
Neutral object vs neutral object	50.92	15.32

Table 1. *Means and standard deviations of percent fixation to the target image for comparisons of all target and distractor combinations.*

4. Discussion

The purpose of the present study was to determine if positive affect could aid infants' acquisition of word-object associations. It was hypothesized that infants would show stronger learning of word-object associations for objects taught with positive affect than those taught with neutral affect. The hypothesis was not supported, as no evidence of word learning was found for objects familiarised in either positive or neutral affect. Infants' did demonstrate looking to the target for familiar objects, indicating that their inability to learn the novel objects was not a failure to perform the task at test. The failure of learning in either condition is in contrast with previous research that finds evidence of associative learning at 17 months [5] and 21 months [6].

Infants in the current study were required to learn four novel word-objects pairs. In contrast, Graf Estes and Hurley [5], taught infants only two novel objects, which is likely to have been less demanding. Further, the two words used in the Graf Estes and Hurley study formed a non-minimal word pair. Thus, infants did not need to encode fine phonological detail in order to succeed in the task. Requiring infants to do so in the present study may have made the task too challenging. However, the current study used a preferential looking task, which is known to be less cognitively demanding [22] than the switch task used by Graf Estes and Hurley.

Similar to the current study, Ma, Golinkoff, Houston, and Hirsh-Pasek [6] used a preferential looking task and infants in their study were able to make word-object associations in IDS at 21 months. However, in Ma et al. [6] as with Graf Estes and Hurley [5], infants were only taught two novel objects, and the two words formed a non-minimal word pair. Further, infants in Ma and colleagues' study were provided with reminder trials halfway through testing, in which the infants were shown each novel object and word pairing again prior to testing trials recommencing. Thus, this study is likely to have been less cognitively demanding for the infants than the current study.

It was further hypothesized that infants would demonstrate preferences for objects familiarised in positive affect over objects familiarised in neutral affect during the behavioural choice test, and there was no evidence to support this hypothesis. This indicates that the affect which was used to label the object did not influence infants' preferences.

The present study revealed that infants demonstrated strong preferences for the familiar images over novel objects. In particular, infants looked more when the target was familiar, both when the distractor was an object familiarised in positive affect and when it was familiarised in neutral affect. This suggests the infants found the familiar objects more interesting than the novel objects, which may provide another explanation for the failure of learning.

Based on the hypothesis that the present study was too cognitively demanding for the infants, a follow-up study is currently underway in our lab. In the new study, 19-month-old infants are exposed to a similar paradigm to the current study; however, only two novel objects and a non-minimal word pair are used, in order to ease the task demands on the infants. Also in this new study, infants' preferences for the novel objects are measured at baseline, with the objects presented side-by-side on the screen without audio prior to testing. As per the current study, the word-object pairings are presented with a clear referential status which has been shown to aid infants' mapping of novel words to objects [21].

In conclusion, the present study demonstrates that infants may be unable to learn word-object associations at 19 months when the task involves four novel word-object pairings comprising two minimal-word pairs. As infants at this age would be expected to be able to make word-object associations, these results suggest the task may have been too cognitively demanding for the infants. Further, infants demonstrate strong preferences for familiar images of a baby and a cat over novel objects and this should be considered when designing future experiments. Ongoing follow-up research aims to mitigate the cognitive demands that have confused the current findings in order to provide more answers on how positive affect is involved in infants' early word learning.

5. Acknowledgments

The authors would like to thank Brandon Huynh, Liquan Liu, Isabel Lopez, Candice Michael, Scott O'Loughlin, Aimee

Oliveri, Valeria Peretokina, Gloria Pino Escobar, Christina Quattropani, and Michelle Williams for their assistance with data collection.

This project was funded by the ARC (Australian Research Council) Centre of Excellence for the Dynamics of Language.

6. References

[1] 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.

[2] A. Fernald and P. Kuhl, ‘Acoustic determinants of infant preference for motherese speech’, *Infant Behav. Dev.*, vol. 10, no. 3, pp. 279–293, Jul. 1987.

[3] J. F. Werker and P. J. McLeod, ‘Infant preference for both male and female infant-directed talk: A developmental study of attentional and affective responsiveness’, *Can. J. Psychol. Can. Psychol.*, vol. 43, no. 2, pp. 230–246, 1989.

[4] L. Singh, J. L. Morgan, and C. T. Best, ‘Infants’ Listening Preferences: Baby Talk or Happy Talk?’, *Infancy*, vol. 3, no. 3, pp. 365–394, Jul. 2002.

[5] K. Graf Estes and K. Hurley, ‘Infant-Directed Prosody Helps Infants Map Sounds to Meanings’, *Infancy*, vol. 18, no. 5, pp. 797–824, Sep. 2013.

[6] W. Ma, R. M. Golinkoff, D. M. Houston, and K. Hirsh-Pasek, ‘Word Learning in Infant- and Adult-Directed Speech’, *Lang. Learn. Dev.*, vol. 7, no. 3, pp. 185–201, Jul. 2011.

[7] A. Fernald and T. Simon, ‘Expanded intonation contours in mothers’ speech to newborns’, *Dev. Psychol.*, vol. 20, no. 1, pp. 104–113, Jan. 1984.

[8] D. L. Grieser and P. K. Kuhl, ‘Maternal speech to infants in a tonal language: Support for universal prosodic features in motherese’, *Dev. Psychol.*, vol. 24, no. 1, pp. 14–20, Jan. 1988.

[9] D. N. Stern, S. Spieker, R. K. Barnett, and K. MacKain, ‘The prosody of maternal speech: infant age and context related changes’, *J. Child Lang.*, vol. 10, no. 1, pp. 1–15, Feb. 1983.

[10] L. J. Trainor and C. M. Austin, ‘Is infant-directed speech prosody a result of the vocal expression of emotion?’, *Psychol. Sci. Wiley-Blackwell*, vol. 11, no. 3, p. 188, May 2000.

[11] H.-M. Liu, P. K. Kuhl, and F.-M. Tsao, ‘An association between mothers’ speech clarity and infants’ speech discrimination skills’, *Dev. Sci.*, vol. 6, no. 3, pp. F1–F10, Jun. 2003.

[12] E. D. Thiessen, E. A. Hill, and J. R. Saffran, ‘Infant-Directed Speech Facilitates Word Segmentation’, *Infancy*, vol. 7, no. 1, pp. 53–71, Jan. 2005.

[13] L. Singh, S. Nestor, C. Parikh, and A. Yull, ‘Influences of Infant-Directed Speech on Early Word Recognition’, *Infancy*, vol. 14, no. 6, pp. 654–666, Nov. 2009.

[14] J. F. Werker, F. Pons, C. Dietrich, S. Kajikawa, L. Fais, and S. Amano, ‘Infant-directed speech supports phonetic category learning in English and Japanese’, *Cognition*, vol. 103, no. 1, pp. 147–162, Apr. 2007.

[15] A. Schachner and E. E. Hannon, ‘Infant-directed speech drives social preferences in 5-month-old infants’, *Dev. Psychol.*, vol. 47, no. 1, pp. 19–25, Jan. 2011.

[16] L. Singh, J. L. Morgan, and K. S. White, ‘Preference and processing: The role of speech affect in early spoken word recognition’, *J. Mem. Lang.*, vol. 51, no. 2, pp. 173–189, Aug. 2004.

[17] L. Singh, ‘Influences of high and low variability on infant word recognition’, *Cognition*, vol. 106, no. 2, pp. 833–870, Feb. 2008.

[18] L. Fenson, V. Marchman, D. J. Thal, P. S. Dale, J. S. Reznick, and E. Bates, *MacArthur-Bates Communicative Development Inventories. User’s guide and technical manual, (2nd ed).*, 2nd ed., vol. 22. Baltimore: Brookes, 2007.

[19] P. S. Kaplan, K. B. Fox, and E. R. Huckeby, ‘Faces as reinforcers: Effects of pairing condition and facial expression’, *Dev. Psychobiol.*, vol. 25, no. 4, pp. 299–312, May 1992.

[20] A. Fernald and N. Hurtado, ‘Names in frames: infants interpret words in sentence frames faster than words in isolation’, *Dev. Sci.*, vol. 9, no. 3, pp. F33–F40, May 2006.

[21] C. T. Fennell and S. R. Waxman, ‘What Paradox? Referential Cues Allow for Infant Use of Phonetic Detail in Word Learning’, *Child Dev.*, vol. 81, no. 5, pp. 1376–1383, Sep. 2010.

[22] K. A. Yoshida, C. T. Fennell, D. Swingley, and J. F. Werker, ‘Fourteen-month-old infants learn similar-sounding words’, *Dev. Sci.*, vol. 12, no. 3, pp. 412–418, May 2009.

Perception of Statement and Question Intonation: Cantonese versus Mandarin

Una Y. Chow, Stephen J. Winters

School of Languages, Linguistics, Literatures and Cultures, University of Calgary, Canada

uchow@ucalgary.ca, swinters@ucalgary.ca

Abstract

This study investigated Cantonese and Mandarin listeners' perception of their native intonation patterns. Twenty native listeners of each language identified the sentence type of gated statements and echo questions. Cantonese listeners were more sensitive to the distinction between statements and questions than Mandarin listeners, even though question intonation cues begin earlier in an utterance in Mandarin than in Cantonese. Both groups were more sensitive to the timing and duration of the F0 cues than to changes in F0 direction, and also performed better when an utterance's final syllable carried low or falling tones, rather than high or rising tones.

Index Terms: intonation, perception, F0, tone, Cantonese, Mandarin, echo question, statement

1. Introduction

Languages differ in their use of intonation to express different types of sentences. English typically signals yes/no questions with a rise in fundamental frequency (F0) at or near the nuclear tone of an intonational phrase [1]. Although many speakers associate the intonation of questions with a high F0 rise [2], some languages do signal questions with a fall in intonation, known as lax prosody [3]. For instance, Mooré, a Gur language, attaches a low boundary tone to a statement to convert it into a question [3].

Even closely related languages like Cantonese and Mandarin exhibit clear differences in the structure of their intonation systems. Cantonese is a tone language with six lexical tones: T1 (high-level [55]), T2 (high-rise [25]), T3 (mid-level [33]), T4 (low-fall [21]), T5 (low-rise [23]), and T6 (low-level [22]). Mandarin is also a tone language, with two fewer lexical tones than Cantonese: T1 (high-level [55]), T2 (rising [35]), T3 (low-fall-rise [214]), and T4 (falling [51]). Cantonese signals echo questions with a high F0 rise at the end of the utterance [4, 5]. Mandarin, on the other hand, signals echo questions with a gradual increase in F0 towards the end of the utterance [6]. In addition, since native Mandarin listeners perceive F0 primarily at the lexical level, lexical tones can reduce their sensitivity to the F0 cues at the intonational level [7].

This study investigated Cantonese and Mandarin listeners' perception of statements and echo questions in their native language in order to determine if it is easier for native listeners to identify statements and echo questions in one language than the other. This study also investigated how lexical tones on the final syllable affected the ability of both groups of listeners to identify these two sentence types in an intonation perception task.

2. Method

A perception experiment was conducted using a modified gating task aimed at finding out how much of the statement-question F0 cues listeners can extract from each gated portion of an utterance.

2.1. Stimuli

Ten native Cantonese speakers (from Hong Kong) and sixteen native Mandarin speakers (from different regions of China, excluding Hong Kong) produced the stimuli for the perception experiment. The speakers, aged 18-35 years, were recruited from the University of Calgary. They reported no visual, speech, or hearing impairments.

The speakers were recorded individually in a sound-attenuated booth. They read five blocks of four dialogues, presented to them one dialogue at a time on a computer screen in Chinese characters. Each dialogue contained a target pair of sentences. The pair consisted of a statement and an echo question that had the same lexical and syntactic form as the statement. A filler question preceded the pair and a filler statement followed it to provide pragmatic context. In each block, the target sentences in the four dialogues ended in the same syllable but with a different tone (e.g., *shi1*, *shi2*, *shi3*, and *shi4*). The target sentences also varied in length across the blocks: 5, 7, 9, 11, and 13 syllables long for blocks A, B, C, D, and E, respectively. All target sentences began with a disyllabic name. (1) shows a sample Mandarin dialogue, in which *Wang1 Wu3 shi4 lao3 shi1* is the target sequence for both the statement and echo question. The Cantonese dialogues were semantically similar to the Mandarin dialogues.

(1) A: 汪五是谁?
Wang1 Wu3 shi4 shei2?
'Who is Wang Wu?'
B: 汪五是老师。
Wang1 Wu3 shi4 lao3 shi1.
'Wang Wu is a teacher.'
A: 汪五是老师?
Wang1 Wu3 shi4 lao3 shi1?
'Wang Wu is a teacher?'
B: 是, 汪五是老师。
Shi4, Wang1 Wu3 shi4 lao3 shi1.
'Yes, Wang Wu is a teacher.'

To create the stimuli for the perception experiment, two male and two female speakers were randomly selected from the production recordings for each language. Their 160 recorded sentences (4 speakers x 20 target pairs x 2 sentence types) were gated in five forms, as shown in Table 1, generating a total of 800 tokens for each language.

Table of Contents for this manuscript

Table 1. *Stimulus types for the perception experiment.*

Stimulus Type	Description	Cantonese Example: ‘Wong Ji (is) not on time’
Whole	Whole sentence	*Wong1 Ji6 m4 zeon2 si4*
NoLast	All but the last syllable	*Wong1 Ji6 m4 zeon2*
Last	Last syllable	*si4*
Last2	Last two syllables	*zeon2 si4*
First	Sentence-initial name	*Wong1 Ji6*

2.2. Listeners

Twenty native Cantonese listeners (seven from Hong Kong, six from Guangdong, and six from Canada) and twenty native Mandarin listeners (from different regions of China, excluding Hong Kong) participated in the perception experiment. Half of the listeners were males and half were females. The listeners, 18-35 years old, were fluent speakers of their native languages. They were recruited from Calgary, Canada and reported no speech or hearing impairments.

2.3. Procedure

The listeners sat in a quiet room and listened to the audio stimuli through headphones. A computer presented the stimuli to the listeners one at a time and also instructed them to rate whether the stimulus they had just heard was a statement or a question by pressing the corresponding key on the keyboard.

The identification task required two testing sessions on two separate days. Session 1 comprised five parts. Parts I-III were practice, training, and testing phases for the stimulus type Whole. Parts IV-V were practice and testing phases for the stimulus types NoLast and Last combined. Session 2 added parts VI-VII, which were practice and testing phases for the stimulus types Last2 and First combined. The reason for excluding parts VI-VII from session 1 was to keep each session within 45 minutes to retain the listener’s attention. In session 1, the listener performed an extra task of completing a language background questionnaire.

This study is part of a larger study that compares human listeners’ performance in the identification task with that of an exemplar-based model in order to find out if exemplar theory can account for the human perception of intonation. Since this model lacks prior language experience and the human listeners also need to be tested in the same way as the model, half of the stimuli were used for training and half for testing. The training and test stimuli were counterbalanced between listeners and between sessions for each listener. In the practice sessions, listeners heard two non-target pairs of statements and echo questions, produced by a different female speaker from the speakers used in testing and training. The practice and training phases provided immediate feedback to the listener, after each trial, on the type of sentence (statement or question) that had been presented in the trial. The testing phases provided feedback after every ten trials only on the number of correct responses in those ten trials, to help keep the listener motivated to perform the task well.

2.4. Analysis

Responses from the perception study were converted, for each listener, into measures of perceptual sensitivity (d') and response bias (ß). In order to compare the effects of the final tone on the listeners’ identification of statements and echo questions between Cantonese and Mandarin, the tones were labeled, grouped, and compared as indicated in Table 2. The reason for grouping the Cantonese tones is that Cantonese has more tones than Mandarin. The rationale for grouping [25] with [23] and [33] with [22] is that previous research has found that it is difficult, even for native speakers, to produce the tones in each pair distinctively [8, 9] and to perceive their differences [10]. These tones could be merging.

Table 2. *Cantonese and Mandarin tonal groups.*

Tonal Group	Cantonese	Mandarin
H (High)	T1 [55]	T1 [55]
R (Rising)	T2 [25], T5 [23]	T2 [35]
L (Low)	T3 [33], T6 [22]	T3 [214]
F (Falling)	T4 [21]	T4 [51]

3. Results

ANOVAs on d' and ß found no significant difference between session 1 and session 2 ($p > .05$) for the stimulus types that were presented in both sessions (Whole, NoLast, and Last). Therefore, only data from session 2 were analyzed, since that set of data contained all five stimulus types.

3.1. Perceptual Sensitivity

Figure 1 shows the Cantonese and Mandarin listeners’ sensitivities to each stimulus type, broken down by tonal group. A three-way ANOVA with d' as the dependent measure and with language, stimulus type, and tone as independent factors found a significant main effect of language [$F(1, 760) = 5.4$, $p = .02$]. There were also significant interactions between language and stimulus type [$F(4, 760) = 35.4$, $p < .001$], between language and tone [$F(3, 760) = 7.1$, $p < .001$], and among language, stimulus type, and tone [$F(12, 760) = 3.0$, $p < .001$]. A post-hoc Tukey HSD test revealed significant differences among these three factors.

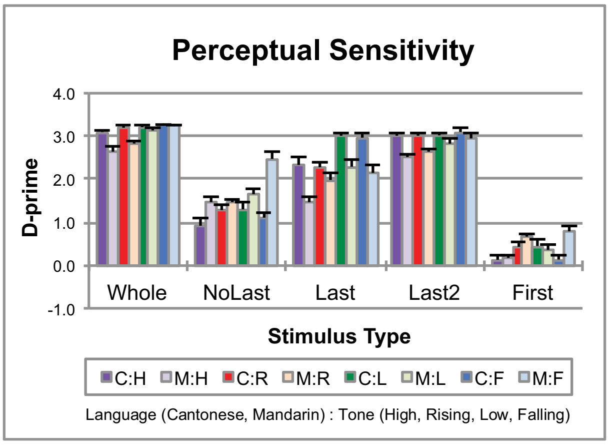


Figure 1: *Cantonese and Mandarin listeners’ sensitivities to each stimulus type, by tonal group.*

Within languages, the Cantonese listeners performed significantly better on stimulus type Last when the final syllable carried a low tone, rather than a high or rising tone ($p < .05$; mean difference ($\bar{d}$) = .66 and .71, respectively), and when the final syllable carried a falling tone, rather than a rising tone ($p < .05$; $\bar{d} = .69$). The Mandarin listeners, on the other hand, performed significantly better on stimulus type Last when the final syllable carried a low or falling tone, rather than a high tone ($p < .05$; $\bar{d} = .84$ and .69, respectively).

They also performed significantly better on stimulus type NoLast when the missing syllable carried a falling tone, rather than a high, rising or low tone (*p* < .001; d̄ = 1.01, 1.02, and .85, respectively).

Between both languages, the Cantonese listeners performed significantly better than the Mandarin listeners in identifying statements and echo questions (*p* = .02; d̄ = .09) on all stimulus types combined. Specifically, on stimulus type NoLast, the Mandarin listeners performed significantly better than the Cantonese listeners only when the missing syllable carried a falling tone (*p* < .001; d̄ = 1.36). However, on stimulus type Last, the Cantonese listeners performed significantly better than the Mandarin listeners when the syllable carried a high, low, or falling tone (*p* < .05; d̄ = .88, .70, and .82, respectively) but not a rising tone.

3.2. Response Bias

A three-way ANOVA with ß as the dependent measure and with language, stimulus type, and tone as independent factors revealed only a significant main effect of stimulus type [*F*(4, 760) = 97.0, *p* < .001] and a significant interaction between language and stimulus type [*F*(4, 760) = 5.1, *p* < .001]. A post-hoc Tukey HSD test revealed that, overall, the listeners showed significantly more bias towards statements on the stimulus types that excluded the final syllable than on the stimulus types that included the final syllable (*p* < .01). Within languages, both the Cantonese and the Mandarin listeners showed significantly more bias towards ‘statement’ responses on stimulus types NoLast and First—neither of which includes the final syllable—than on stimulus types Whole, Last, and Last2 (*p* < .001). The Cantonese listeners also showed significantly more bias towards statements on stimulus type NoLast than stimulus type First (*p* = .04; d̄ = .21) and on stimulus type Whole than stimulus type Last (*p* = .02; d̄ = .23). Between both languages, Mandarin showed significantly more bias towards ‘statement’ responses on stimulus type First than the Cantonese listeners (*p* = .003; d̄ = .26).

3.3. F0 Analysis of the Stimuli

To determine if the listeners’ performance on the identification task depended on the F0 contours of the statements and questions, the F0 values at eleven equidistant time points in the stimuli were extracted. Figure 2 shows the mean F0 of each time point of the statement (S) and question (Q) contours by stimulus type. For each language and stimulus type, a two-way repeated measures ANOVA was performed with F0 as the dependent measure and with sentence type and time point as the independent factors. On stimulus type Whole, the question contours showed significantly higher mean F0s than the statement contours at time point 10 for Cantonese [*F*(10, 30) = 4.0, *p* < .001] and at time points 8-10 for Mandarin [*F*(10, 30) = 26.1, *p* < .001]. On stimulus type NoLast, the question contours showed significantly higher mean F0s than the statement contours at time points 8-10 for Mandarin [*F*(10, 30) = 12.2, *p* < .001]. There was no significant difference on stimulus type NoLast for Cantonese (*p* > .96). On stimulus type Last, the question contours showed significantly higher mean F0s than the statement contours at time points 5-10 for Cantonese [*F*(10, 30) = 23.55, *p* < .001] and at all eleven time points for Mandarin [*F*(10, 30) = 27.0, *p* < .001]. On stimulus type Last2, the question contours showed significantly higher mean F0s than the statement contours at time points 8-10 for Cantonese [*F*(10, 30) = 27.0, *p* < .001] and, again, at all eleven time points for Mandarin [*F*(10, 30) = 19.9, *p* < .001]. On stimulus type First, there was no significant difference for either language (*p* > .57).

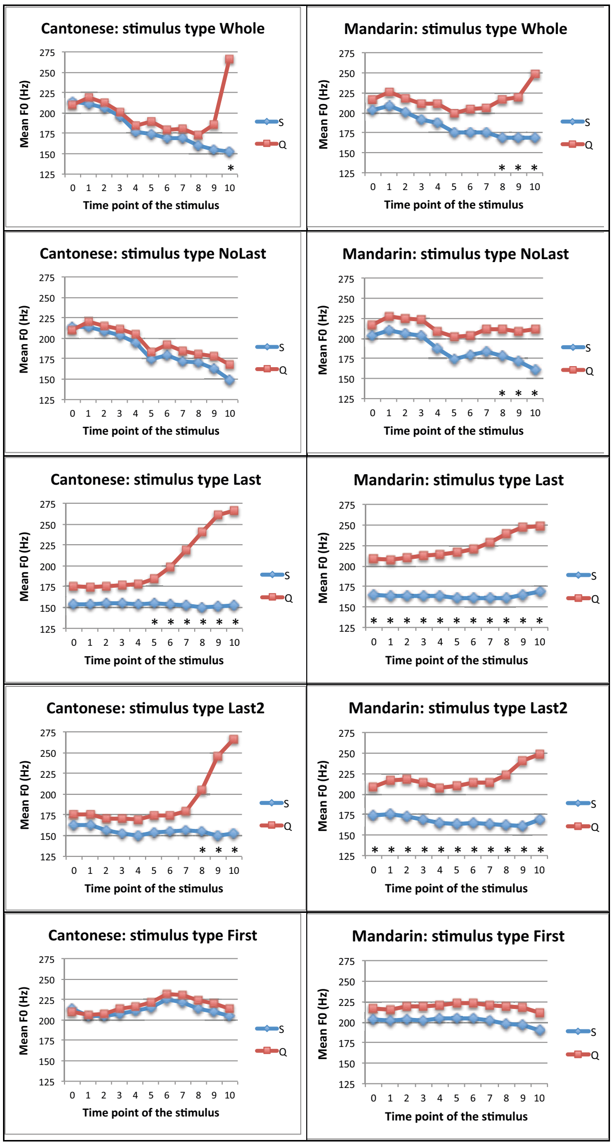


Figure 2: *Significant differences (* p < .05) between statements (S) and questions (Q) at time points 0-10 of all five stimulus types for Cantonese and Mandarin.*

A repeated measures ANOVA was also performed for each tonal group of the stimulus type, with F0 as the dependent measure and with sentence type and time point as independent factors. For Cantonese, the significant results of the tonal groups were similar to the overall result of each stimulus type. For Mandarin, however, the question contour had significantly higher mean F0s than the statement contour at time points 7-10 when stimulus type Last2 ended in a high tone [*F*(10, 30) = 18.9, *p* < .001], at time points 9-10 when stimulus type Last2 ended in a rising tone [*F*(10, 30) = 20.0, *p* < .001], at time points 9-10 when stimulus type Last2 ended in a low tone [*F*(10, 30) = 11.3, *p* < .001], and at time points 1-10 when stimulus type Last2 ended in a falling tone [*F*(10, 30) = 10.2, *p* < .001].

4. Discussion

The echo question cue in Mandarin starts earlier in the utterance, usually before the final syllable, whereas the echo question cue in Cantonese starts in the final syllable. This accounts for why the Cantonese listeners performed better than the Mandarin listeners when presented with just the final syllable, while the Mandarin listeners performed better than the Cantonese listeners on utterances excluding just the final syllable. In addition, the Cantonese listeners performed relatively better on all except the rising tones because the rising tone's F0 pattern is confusable with the question rise [11]. Furthermore, researchers have reported that Mandarin listeners identify questions ending in the falling tone much more accurately than questions ending in the other three tones [12], so it seems odd that the Mandarin listeners in this study performed better when the excluded final syllable on the utterance carried a falling tone, rather than any of the other three tones. This happened because the F0 gap between the statement and question intonation contours at time point 0 of the final syllable (which is time point 10 of stimulus type NoLast) is greater for the falling tone than the high, rising, and low tones, as Figure 3 shows. Therefore, although Cantonese's question rise enabled the Cantonese listeners to perform better than the Mandarin listeners in the identification task overall, a final rising tone in statements reduced the Cantonese listeners' sensitivities to the F0 cues in question intonations and lowered their performance on stimulus type Last, as shown in Figure 1. On the other hand, a final falling tone in Mandarin questions increased the Mandarin listeners' sensitivities to the F0 difference between the statement and question intonations on the non-final portions of utterances ending in the falling tone.

Figure 3 also shows that the F0 direction of Cantonese statements ending in a high tone is falling at time points 6-8—similar to the ending for the low tone—but the Cantonese listeners performed better on the low tone than the high tone. Similarly, the F0 directions of Mandarin questions ending in the low and falling tones at time points 5-10 are rising and falling, respectively, but the Mandarin listeners performed better on the low and falling tones than on the high tone. This suggests that listeners may have based their responses more on the F0 difference between the statement and echo question, as well as on the timing of this difference in the final syllable, than on the patterns in F0 direction.

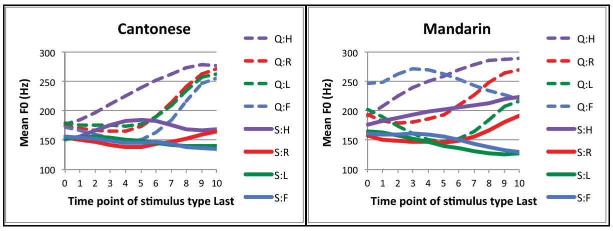


Figure 3: *Tonal contours of the question (Q) and statement (S) stimuli Last for Cantonese and Mandarin.*

Listeners tend to be biased towards the perception of statements [12, 13] when they cannot distinguish them from questions. Similar to the Cantonese listeners, the Mandarin listeners showed more bias towards 'statement' responses on stimuli that lack the final syllable than on stimuli that contain the final syllable. This suggests that, for Mandarin, the most salient F0 cue is in the last syllable even though other cues exist elsewhere in the utterance. The acoustic analysis of stimulus type First did not reveal any significant F0 difference between the statement and echo question contours. However, the Mandarin listeners showed significantly more bias towards statements for these stimuli than the Cantonese listeners. Perhaps the Mandarin listeners were expecting a 'raised pitch range' [14] at the onset of the question. Since this cue was missing, they misperceived the stimulus as a statement.

5. Conclusions

Overall, the Cantonese listeners in this study identified statements and questions in their native language more accurately than the Mandarin listeners. This result suggests that a shorter cue with a greater F0 gap between the statement and echo question contours may be easier to detect. However, a longer cue with a smaller F0 gap between the two sentence-type contours may serve as a more robust cue in situations where part of the cue may not be heard.

6. Acknowledgements

This research was supported by the Social Sciences and Humanities Research Council of Canada.

7. References

[1] Wells, J. C., English Intonation: An Introduction, Cambridge University Press, 2006.

[2] Gussenhoven, C. and Chen, A.-J., "Universal and language-specific effects in the perception of question intonation", Proc. Spoken Language Processing, 91-94, 2000.

[3] Rialland, A., "The African lax question prosody: Its realisation and geographical distribution", Lingua, 119(6):928-949, 2009.

[4] Gu, W., Hirose, K. and Fujisaki, H., "Analysis of the effects of word emphasis and echo questions on F0 contours of Cantonese utterances", Interspeech, 1825-1828, 2005.

[5] Wong, W. Y. P., Chan, M. K. M. and Beckman, M. E., "An autosegmental-metrical analysis and prosodic annotation conventions for Cantonese", in S.-A. Jun [Ed], Prosodic Typology: The Phonology of Intonation and Phrasing, Oxford University Press, 271-300, 2005.

[6] Liu, F., Surendran, D. and Xu, Y., "Classification of statement and question intonations in Mandarin", Proc. Speech Prosody, 2006.

[7] Yuan, J., "Perception of intonation in Mandarin Chinese", J. Acoust. Soc. Am., 130(6):4063-4069, 2011.

[8] Bauer, R. S. and Benedict, P. K., Modern Cantonese Phonology. Trends in Linguistics Studies and Monographs 102, Mouton de Gruyter, 1997.

[9] Mok, P. P. K., Zuo, D. and Wong, P. W. Y., "Production and perception of a sound change in progress: Tone merging in Hong Kong Cantonese", Language Variation and Change, 25:341-370, 2013.

[10] Ciocca, V. and Lui, J., "The development of lexical tone perception in Cantonese", J. Multilingual Communication Disorders, 1:141-147, 2003.

[11] Ma, J. K.-Y., Ciocca, V. and Whitehill, T., "Effect of intonation on Cantonese lexical tones", J. Acoust. Soc. Am., 120:3978-3987, 2006.

[12] Yuan, J.-H., "Perception of Mandarin intonation", in Proc. International Symposium on Chinese Spoken Language Processing, 45-48, 2004.

[13] Ma, J. K.-Y., Ciocca,V. and Whitehill, T. L., "The perception of intonation questions and statements in Cantonese", J. Acoust. Soc. Am., 129(2):1012-1023, 2011.

[14] Peng, S.-H., Chan, M. K. M., Tseng, C.-Y., Huang, T., Lee, O. J. and Beckman, M. E., "Towards a Pan-Mandarin system for prosodic transcription", in S.-A. Jun [Ed], Prosodic Typology: The Phonology of Intonation and Phrasing, 230-270, Oxford University Press, 2005.

Prosodic characteristics of Japanese polite speech spoken by native and non-native speakers

Chiharu Tsurutani[1], *Shi Shuju*[2], *Nobuaki Minematsu*[2]

[1] School of HLSS, Griffith University, Australia
[2] Dept. of Information and Communication Engineering, The University of Tokyo, Japan
C.Tsurutani@griffith.edu.au, {shishuju, mine}@gavo.t.u-tokyo.ac.jp

Abstract

We conducted experiments in the production and the perception of Japanese polite speech to explore the effectiveness of prosodic features on the perception of politeness. We recorded common polite forms, in polite and non-polite scenarios using 10 native and 10 non-native speakers, and 10 listeners assessed their level of politeness. The results showed that: 1) native speakers tend to use a lower pitch register, a narrower pitch range and a slower speech rate, 2) there is a certain type of pitch modification observed in polite speech; 3) learners seemed to use only a slower speech rate to express politeness.

Index Terms: polite speech, prosodic features, Japanese, native and non-native speech

1. Introduction

Prosody plays an important role not only in intelligibility but also in speakers' attitudes and emotions. It is a general consensus that intonation can completely change the interpretation of the utterance. Prosodic characteristics in speech conveying emotion and attitude share features across all languages. Ohala [1] argues that the use of high pitch conveys an impression of subordination and submissiveness and is considered to be non-threatening, conveying goodwill. High pitch also leads to high scores in polite, non-aggressive, friendly and other positive attitudes in perception studies [2] [3]. Nevertheless, attitudes associated with social behavior can have language specific prosodic characteristics.

In this study, politeness was chosen to investigate possible language specific requirements in Japanese speech. Japanese people are known to be polite and well-mannered in public. Learning to sound polite is particularly important for second language learners (L2) of Japanese. As Japanese has elaborate honorific systems, politeness studies and guidance for L2 learners tend to focus on morphological and lexical aspects, but the information on appropriate intonation for polite speech has not been made available to L2 learners. This study attempts to identify common prosodic features used by native speakers in polite speech and to provide some strategies useful in delivering polite speech for L2 learners of Japanese. To this end, the following steps were taken, which are used in structuring the paper: After reviewing the phonetic aspects of politeness dealt with in previous studies, an experiment was conducted to collect polite and non-polite speech from both native and non-native speakers. The politeness of their performance was judged by native listeners, and statistically analyzed to identify factors that affect the degree of politeness.

2. Phonetic aspects of politeness

The degree of politeness can be affected by non-linguistic factors such as non-verbal expressions, the contents and context of the utterance. However, after considering all these factors, there are still some phonetic characteristics which can differentiate between polite and non-polite utterances. Brown et al. [4] investigated the significance of phonetic information in Korean without morphological and lexical marking. The perceptual ability of native and non-native listeners who had no knowledge of Korean was tested on a sentence inserted in casual and honorific conversation, respectively. Both native and non-native speakers were able to identify polite speech above chance level, which implies the existence of some cross-linguistic cues for vocal politeness. They used similar cues such as pitch and vowel quality, but the ranking and importance of cues differed between the two groups, which indicates there are also features that are specific to certain languages and cultures. There are several prosodic factors that have been commonly investigated for their influence in expressing politeness in previous studies. These are discussed in the following sections.

2.1. Pitch register (average pitch)

As a general impression, Japanese female speakers are known to use high pitch for polite speech [5] [6]. However, high pitched speech is not always observed in polite speech in all language communities. Grawunder and Winter [7] found that polite speech in Korean had a lower pitch level and narrower pitch range than informal speech. Recently Idemaru, Winter and Brown [8] added new information concerning pitch register in Korean polite speech in terms of perception. They found that female Korean listeners perceived high pitch as being polite, while male Korean listeners perceived it to be informal. Their study used informants (6 females and 4 males) who resided in the US, including one male listener who was born in the US. Although it is difficult to draw a conclusion from this one perception study, it does imply that pitch register is not an absolute determiner of politeness.

2.2. Pitch range

In the study of attitudinal prosody in Mandarin Chinese, it was reported that polite speech had wider pitch, as well as higher register of pitch, than neutral speech [9]. On the other hand, Grawunder and Winter's study on Korean language reported a narrower pitch range for polite speech. The different result can be attributed to different prosodic requirements between the two languages; a tone language vs. a non-tonal language. Or it could be due to the sentence type used as stimuli, whether it is a question or a statement. Ofuka et al. [10] investigated six Japanese male speakers' production of polite speech and found that both average pitch and pitch range did not become cues for politeness. In another study of Japanese polite speech, increase of pitch range in casual speech addressed to intimate friends was observed in the interview to the participants [5]. Using both male and female speakers and various sentence types, pitch range in Japanese polite speech needs to be reexamined further.

2.3. Speech rate

Speech rate has been regularly discussed in polite speech [5] [7] [10]. The general impression that slow speech sounds more polite has been proven in studies of Korean speech. On the other hand, studies into Chinese and Japanese speech reports otherwise. Wentao et al. [9] reported that polite speech in Chinese was faster. Ofuka et al. [10] did not find any conclusive tendency, apart from listeners' preference for a speech rate similar to their own, even though the rate factor showed some relevance to politeness. Clearly, speech rate is a factor to be investigated. In this study, speech rate will be measured in the form of the duration of the utterance, since it is known that listeners gain an impression of slow speech by listening to the entire speech [11].

2.4. F0 direction and duration of the final syllable of the sentence

In previous studies on Japanese politeness, prosodic information of final particles was paid attention to as a cue for politeness. Japanese has a rich inventory of sentence final particles. Interrogative sentences are expressed with "*ka*" without changing the word order from a declarative sentence. The intonation is normally raised for question sentences, as is the case for over 70% of the world's languages [12]. The question sentence without "*ka*" will require a rise in the sentence's final tone more than in unmarked question sentences, which will be the case in other languages as well. The problem is that the pitch rise in the sentence ending with the final particle "*ka*" does not always occur in actual speech. In Ofuka, et al.'s study [10], 2 out of 6 speakers had a fall in the F0 direction in the final vowel, and the proportion of rise and fall was half and half in Ogino and Hong's data [5]. However, in their politeness judgement test [10], native listeners preferred the unmarked sentence intonation, a final pitch rise, to a final fall. They further investigated the prosody of the final vowel of the sentence and found a great impact on politeness judgements in duration and pitch change, in particular in the last 100ms. The sentence's final prosody is considered to accurately convey the speaker's attitude or intention. In the experiment of this study, pitch register, pitch range, speech rate and the prosodic information in the final syllable of the sentence will be investigated.

3. Experiment

Stimulus sentences were prepared considering the following points: speech style, sentence type, and the type of attitudinal speech. *Desu/masu* form was chosen as it is a common polite form familiar to L2 learners at all levels of proficiency. In previous studies, only one sentence type was used, which was mostly a question [5] [10]. Whether the sentence is a question or a statement could influence pitch height and pitch range. Thus, both types of sentences were included. From the attitudinal speech categories listed in Japanese textbooks, Request, Inquiry and Decline were selected, as they could present a different level of politeness by prosody more clearly than other functions.

3.1. Materials

Morphologically and lexically identical sentences, which are crucial for acoustic comparison, were used for polite and non-polite expressions. The target sentences followed the phrases which created either a polite setting or a non-polite setting. In the table below, the clear column, (a), shows polite versions and the shadowed column, (b), has non-polite versions.

Table 1. *Stimulus sentences*

	First phrase to extract the settings	Target sentences
1	a. *Kachoo* (Boss),	*Konoken ni tsuite, iken o onegai dekimasuka* (Can I ask your opinion about this matter?)
	b. *Dareka* (Someone),	
2	a. *Ie, mada Download shiteimasennode* (No, I have not downloaded it yet).	*Sonomamani shiteoite kudasai* (Please leave it as it is.)
	b. *A, hakuban wa kesanaide* (Don't wipe the whiteboard.)	
3	a. *Ie, jitsuwa.* (Well, actually)	*Shiryooga mitukaranai ndesuga, dokode sagashitara iideshooka.* (I cannot find the material, where can I find it?)
	b. *Kono site wa totemo wakarinikui desune.* (This site is hard to understand.)	
4	a. *Soodesune*	*Chotto yoteio mite mimasuga, muzukashiito omoimasu* (I will check my schedule, but think it is difficult)
	b. *Soodesune* Let me see.	
5	a. *Mooshiwakearimasen*	*Mata, jikaini onegai dekimasuka* (Could you please try again next time?)
	b.*Mooshiwakearimasen* (I am sorry.)	
6	a. *Shusseki shitakatta nodesuga, zannendesu.* (Shame, I cannot attend.)	*Tsuginokikaini shussekisasetekudasai.* (Please let me attend next time.)
	b. *Sumimasen* (I am sorry.)	

3.2. Participants

For a small payment, 10 native (5M, 5F) and 10 non-native (6M, 4F) speakers participated in the recording. Native speakers who speak the standard Tokyo dialect were recruited. Their age ranged from 29 to 63. All of them had spent a substantial period of time as office workers at the time of recording. Non-native speakers were students who had been studying in Japan to complete a one year exchange program. Their level of Japanese was intermediate and they were aware of the existence of Japanese polite speech. The utterances in the laboratory recording heavily rely on the participants' role-playing ability. However, in order to extract the same sentence pattern between polite and non-polite versions, we had to rely on a controlled setting, and prepared a scenario which required a polite or non-polite utterance. More than 10 participants were recruited in both groups. The recordings of the 10 best speakers from both groups whose speech clearly exhibited difference between polite and non-polite versions were chosen and used as stimuli.

3.3. Procedure

3.3.1. Methods of recording
The script was presented on the computer screen with the description of the situation and appropriate photos. The researcher operated the computer screen and reminded the speaker of the context of each utterance before s/he produced the sentence. The participants were asked to read out the stimuli sentences twice in the context described on the screen, imagining themselves in the described situation. The participants were told that the first half (1a~6a) of the list required them to be polite towards their senior, while the second half (1b~6b) required them to act as a person in a higher social position so that they could be firm and blunt.
3.3.2. Perception test
10 native listeners (5M, 5F), recruited separately from speakers of the recording, but in a similar age range participated in the perception test. The stimuli were presented to the subjects as an online perception task which was created using Praat. The subjects were asked to score the politeness of the stimulus utterance on a 5-point Likert scale as follows: <5. Very polite ~1. Not at all polite>. There were three exercise files to let the participants familiarize themselves with the task. They were allowed to listen to the stimuli for up to three times in each trial. 6 sentences were divided into two blocks, Block 1 (Short sentences 1,3,5) and Block 2 (Long sentences 2,4,6), depending on the length of the sentences, so that the length does not affect their judgement. Polite and non-polite speech by both native and non-native speakers were mixed and presented in a randomized order.

4. Results

The politeness score for native and non-native speakers' polite and non-polite speech were calculated below in Figure 1.
We evaluated speech to be polite when the score was above 3.5 (dotted line in the graph)

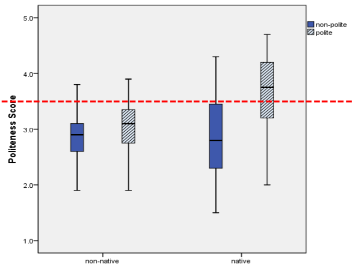
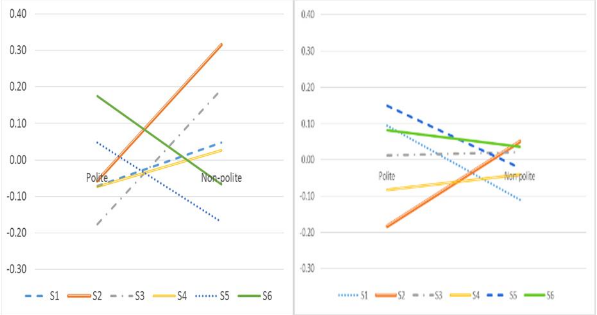


Figure 1: *Politeness scores for native and non-native speakers*

The dark lines in the boxes indicate the mean scores for each category. As found in previous studies, the majority (70% in this study) of polite speech by native speakers was judged as being polite, while only 20% of polite speech by non-native speakers was judged as being polite. We would like to find out what kind of prosodic features contribute to this difference. In the following section, the acoustic measurement for 6 sentences between native and non-native speakers is compared.

4.1. Pitch register

A Z-score was used to normalize the difference between male and female voices. The scores from the 6 sentences were calculated separately. The dotted lines are questions ending with "*ka*" and the solid lines are statements and polite imperatives (please do). The left and right ends of the bars in the graph indicate the scores for polite and non-polite speech respectively.

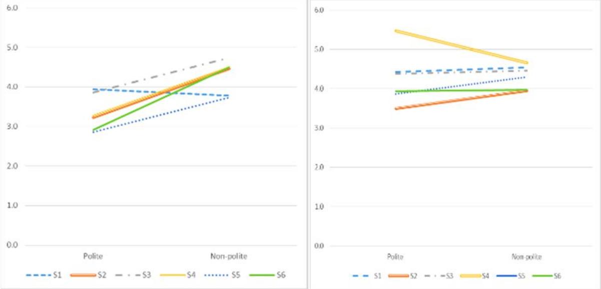


Native speakers Non-native speakers
Figure 2: *Average pitch of 10 speakers (Z-score)*

In native speakers' production, a lower pitch was used for polite speech in 4 sentences, and Sentences 5 and 6 show the opposite pattern. In this data, however, statistical significance was not observed. The general idea that polite speech has a high pitch was not supported by this group of native speakers. No particular pattern was found in non-native speakers' production.

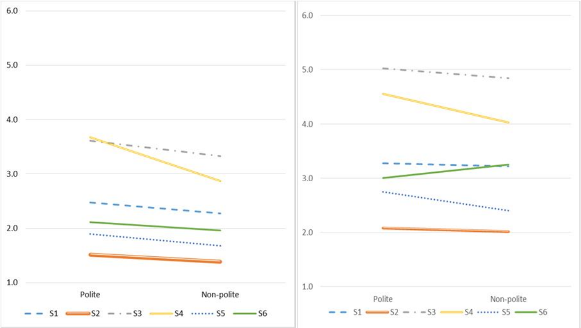
4.2. Pitch range

In the graph depicting the native speakers, the polite versions have a narrower pitch range than the non-polite versions in most sentences. The difference between two versions was statistically significant ($t(59)=5.497$, $p=0.00$). On the other hand, non-native speakers did not show any consistent patterns.



Native speakers Non-native speakers
Figure 3: *Pitch range (Z-score)*

4.3. Speech rate



Native speakers Non-native speakers
Figure 4: *Duration of speech (including pauses)*

It has been reported that listeners perceive the speech rate as an overall impression of a sentence including pauses. The duration of the speech instead of the speech rate was used as an index to measure the perception of slow speech. We can see that slower speech is perceived as being more polite than faster speech ($t(59)$ = -6.067, p=0.00). This is the same result as found in a study of Korean polite speech by native speakers [7]. It would be interesting to find out whether this comes from cultural similarities between the two languages.

4.4. F0 direction and duration of the final syllable of the sentence

The acoustic measurement in sentence final prosody was looked at in comparison with politeness scores, together with the acoustic measurement of the entire utterance.

4.4.1. *The pitch movement in the final particle*

There were three sentence final endings, "*ka", "sai", "mas"*. In "*sai", "mas",* pitch was mostly lowered. In the case of the Wh-question sentence, Sentence 3, "*ka"* has a definite pitch fall. "*ka*" in yes-no questions, Sentences 1 and 5 showed a contrastive difference between native and non-native speech. The pitch pattern, LH was frequently observed in native speakers' production as shown below;

Table 2: *Pitch movement in "ka"*

Pitch pattern	Native	Non-native
LH	11	5
H	3	6
L	4	9

The use of LH in "*ka*" instead of a simple rise must have given a polite impression, as it was observed in many utterances with high scores, which inevitably made the final syllable longer (see Table 3). Native speakers' *ka* is mostly of 2 mora length, while non-natives' "*ka"* was generally shorter than two mora.

Table 3: *Duration of final particle "ka" in Sentence 1 (Duration = Number of mora,* Score=politeness score)

Natives				Non-natives			
Score	Gen	No. mora	Pitch pattern	Score	Gen	No. mora	Pitch pattern
4.7	F	2.0	H	3.8	M	1.6	H
4.7	F	1.9	LH	3.7	M	1.1	H
4.2	F	2.4	LH	3.4	F	2.5	LH
4.0	M	1.9	LH	3.4	M	2.0	L
3.9	F	2.6	LH	3.3	M	1.5	L
3.9	M	2.4	LH	3.3	M	1.5	H
3.5	F	2.2	H	3.1	F	2.0	H
2.8	M	1.7	L	2.9	F	1.5	L
2.7	M	2.7	LH	2.5	F	1.4	LH
2.2	M	1.8	H	2.5	M	1.8	L

4.4.2. *Delay of pitch rise*

In several places in native speakers' polite speech, a delay of pitch rise after initial lowering was observed. This change of pitch pattern was followed by LH in "*ka*" and seems to contribute to a narrowing pitch range in polite speech.

e.g. Sentences 1/5 (correct pattern)
Onegai dekimasu ka (LHHH HHHL H?)
➔ LLHH LLHL LH?

The correlation coefficient with politeness scores for all acoustic measurements was also calculated. In native speakers' production, duration of utterance had the highest correlation (r=0.557), followed by duration of sentence final syllable (r= 0.346) and pitch range (r=-0.232). In the non-native speech, only duration of utterance showed a correlation with politeness, scoring (r=0.212). This means that polite speech was longer, but pitch range was not effectively changed to express politeness in non-native production

5. Conclusions

Prosodic characteristics of Japanese polite speech found in this study are summarized in the following five points; (1) Polite speech does not always have high pitch. Although there was no statistically significant difference, a lower pitch register than non-polite speech was often observed in native speech. (2) Pitch range was narrowed in most polite speech, as found in Korean polite speech. Japanese polite speech could be prosodically very close to Korean polite speech. (3) Slower speech was perceived as being polite. (4) LH Pitch rise at the end of yes-no question sentences was often observed. (5) Delay of pitch rise was observed in the second mora of the word with flat pitch pattern. Non-native speakers only had the characteristic (3) in their polite speech, which means that the other characteristics will be the points for them to pay attention to. The lack of these characteristics is probably the reason for their politeness scores being lower than those of native speakers. Further study is required to identify prosodic factors which affect the impression of the politeness.

6. Acknowledgement

This study was funded by HAKUHO Foundation fellowship, Japan.

7. References

[1] Ohala, J.J. (1984). An ethological perspective on common cross-language utilization of F_0 in voice. *Phonetica*, 41, 1-16.

[2] Nadeu, M., & Prieto, P. (2011). Pitch range, gestural information, and perceived politeness in Catalan. *Pragmatics*, 43(3), 841-854.

[3] Menzes, C., Erickson, D., & Franks, C. (2010). Comparison between linguistic and affective perception of sad and happy – A cross-linguistic study. *Proc. Speech Prosody*, Chicago, IL.

[4] Brown, L., Winter, B., Idemaru, K. & Grawunder, S. (2014). Phonetics and politeness: Perceiving Korean honorific and non-honorific speech through phonetic cues, *J.of Pragmatics Vol. 66, 45-60.*

[5] Ogino, T, and Hong, M. (1992). Nihongo onsei no teineisa ni kansuru kenkyuu (A study on politeness in Japanese speech) In Kunihoro, T. (ed.) The state of the art and analysis of Japanese intonation, Ministry of Education, Tokyo, 215-258.

[6] Ohara, Y. (2001) Finding one's voice in Japanese: a study of the pitch levels of L2 users, In A. Pavienko,, A. Brackledge, I. Piller, M. Teutsch-Dwyer (Eds.), Multilingualism, Second Language Learning, and Gender. Mouton de Gruyter, New York, 231-254.

[7] Grawunder, S., & Winter, B. (2010). Acoustic correlates of politeness: prosodic and voice quality measures in polite and informal speech of Korean and German speakers. *Proc. Speech Prosody*, Chicago, IL.

[8] Idemaru, K., Winter, B. and Brown, L. (2015) The role of pitch in perceiving politeness in Korean, ICPhS.

[9] Gu, W., Zhang, T., & Fujisaki, H. (2011). Prosodic analysis and perception of Mandarin utterances conveying attitudes, *Proc. INTERSPEECH*, Florence, Italy, 1069-1072.

[10] Ofuka, E., McKeown, D., Waterman, M., & Roarch, P. (2000). Prosodic cues for rated politeness in Japanese speech. *Speech Communication*, 32, 199-217.

[11] Kagomiya, T. Yamazumi, Maki, Y. and Maekawa, K. (2008) Factors that affect global perceived speaking rate of spontaneous speech *Journal of* the *Phonetic Society of Japan*, 12 (1), 54-62.

[12] Bolinger, D. L. (1978). Intonation across languages, In J.P. Greenberg, C.A. Ferguson, and E.A. Moravcsik (eds.), *Universals of human language*. Vol.2 *Phonology*. Stanford University Press.

Accentual lengthening in 5-year-old AusE-speaking children: preliminary results

Ivan Yuen, Nan Xu Rattanasone, Elaine Schmidt, Gretel Macdonald, Rebecca Holt, Katherine Demuth

Department of Linguistics, ARC Centre of Excellence in Cognition and its Disorders, Macquarie University

ivan.yuen@mq.edu.au;nan.xu@mq.edu.au;aeis2@cam.ac.uk;gretel.macdonald@gmail.com;
rebecca.holt@mq.edu.au;katherine.demuth@mq.edu.au

Abstract

Although contrastive focus is reported in children's productions, their perception remains poor. However, judgement of contrastive focus in production is typically based on perceptual evaluation, and the stimuli in production and perception studies often differ in the number of syllables and sentence positions. Since both factors influence accentual lengthening in adults, this raises questions about children's ability to produce and generalize the durational cue to focus across different items. Eight AusE adults and 8 children participated in an elicited production task. Unlike the adults, children used accentual lengthening only on monosyllabic, not disyllabic words, and showed no additional phrase-final lengthening.

Index Terms: accentual lengthening, phrase-final lengthening, Australian English, contrastive focus

1. Introduction

When prosody is used to emphasize a specific word in an utterance, this function is known as 'focus' or 'accentuation'. A focused word is typically pitch accented, with increased pitch, duration and intensity [1], [2], [3], [4]. However, children show a discrepancy between their production and perception of focus [5]. Children's ability to produce appropriate adult-like acoustic cues to signal focus in previous studies might have been over-estimated. As the target words in previous production and perception studies varied not only in the number of syllables, but also sentence positions, the discrepancy might have resulted from children's inability to generalize cues to focus across number of syllables and sentence positions in production and perception.

In the adult literature, there is evidence that the number of syllables within a word influences the magnitude of accentual lengthening, one of the cues to signal focus. For instance, [6] found that Scottish English-speaking adults exhibited more lengthening on the accented syllable of a monosyllabic word (e.g., bake enforce) than a disyllabic word (e.g., bacon force). In addition, it is also well-documented that words in sentence-final position undergo lengthening and this could interact with number of syllables [7], and accentual lengthening [6].

In examining American English-speaking children's production of contrastive focus, [8] found that the 4-year-olds could use contrastive focus to indicate a new element in a picture description task. Children were shown a pair of pictures which differed only in one element (e.g., a girl petting a cat vs. a girl petting a dog). While the children did produce contrastive focus, it occurred mostly on the subject noun phrase, and less on the verb or the object noun phrase. This usage pattern, however, could be related to sentence positions, since the subject noun phrase occurs sentence-initially and the object noun phrase appears sentence-finally. Given that children as young as 2 years have been shown to use phrase-final lengthening [9], perhaps children at 4 might have confused phrase-final lengthening with accentual lengthening on the object noun phrase in the sentence-final position. As a result, it might have been difficult to detect the sentence-final contrastive focus through perceptual evaluation of the data.

In addition, [10] observed that children aged 5 made more production errors for the non-final focus items in PEPS-C, such as 'I want a GREEN car.' than for the final focus items, such as 'I want a green CAR.' There were also many ambiguous responses in children's productions of monosyllabic focused words in sentence-final position, with children showing a strong trend of not emphasizing utterance-final words even in obligatory contexts. Yet these data were also perceptually scored. It is then not clear if children can produce appropriate acoustic cues to indicate contrastive focus, specifically the use of accentual lengthening as separate from phrase-final lengthening.

In evaluating children's production of contrastive focus on monosyllabic words (e.g. show bob a BOT), American English speaking adults performed worse on items produced by 4-year-olds (50%) than by 7- and 11-year-olds (above 80%) [11]. Adult listeners also appeared to be less successful at identifying contrastive focus in the non-final than the final position. This could also be related to the acoustic cues that these three groups of children used. The 4-year-olds primarily used duration to indicate contrastive focus, the 7-year-olds used f0, and the 11-year-olds used both f0 and duration to do so.

These studies then raise some questions about children's ability to produce an appropriate durational cue to contrastive focus at age 5 when their focus productions are still ambiguous, especially in relation to their use of accentual lengthening on items with different number of syllables.in different sentence positions.

1.1. Predictions

We tested 5-year-old AusE-speaking children and compared their productions to an adult baseline. H_1: we predicted that adults would exhibit accentual lengthening for both mono- and disyllabic words. H_2: we expected children to show accentual lengthening for monosyllabic words, specifically in phrase-final position. If accentual lengthening is a robust cue to focus, we also expected children to generalize accentual lengthening to disyllabic words. H_3: we also predicted that the 5-year-olds and the adults would implement phrase-final lengthening on focused words in sentence-final position to a greater extent than in sentence-medial position if phrase-final lengthening is separate from accentual lengthening.

2. Method

2.1. Participants

Eight (1M, 7F) monolingual Australian English (AusE) speaking adults were recruited (Age range: 18 - 30 years; Mean = 19;10 years). All were undergraduates at Macquarie University, Sydney, and participated for course credit. Eight monolingual AusE-speaking children (3 M, 5 F) were recruited from the Sydney area, ranging in age from 5;1 to 6;9 years (Mean = 5;8 years).

2.2. Stimuli

The stimuli consisted of an adjective (a colour term) and a noun, forming a noun phrase, for example, '*green ball*'. A set of four adjectives (i.e. *green, grey, orange, yellow*) and a set of eight nouns (i.e. *ball, doll, moon, shoe, bottle, button, pencil, table*) were chosen to generate two types of stimuli: four disyllabic noun phrases (e.g., *green ball*), and four quadri-syllabic noun phrases (e.g., *orange bottle*) to allow us to examine the realization of accentual lengthening across different numbers of syllables. Eight noun phrases were formed to serve as stimuli. Focus location was manipulated to fall either on the adjective or on the noun within the stimuli. These stimuli were embedded in a carrier sentence 'I have a/an X'. Since the adjective must precede the noun of the noun phrase syntactically, the focused adjective ended up being in sentence-medial position and the noun in sentence-final position. To disentangle the potential confound of sentence position from accentual lengthening, we added another condition in which we embedded the focused nouns in the longer carrier sentence 'I have a/an X now'. As a result, there were three experimental conditions, with 8 stimuli in each. This yielded a total of 24 stimuli (Table 1).

Table 1. *Stimuli.*

ADJ-FOC	*I have a GREEN moon/shoe*
	I have a GREY ball/doll
	I have an ORANGE button/table
	I have a YELLOW bottle/pencil
FINAL	
N-FOC	*I have a green MOON/SHOE*
	I have a grey BALL/DOLL
	I have an orange BOTTLE/PENCIL
	I have a yellow BUTTON/TABLE
NON-FINAL	
N-FOC-now	*I have a green MOON/DOLL now*
	I have a grey BALL/SHOE now
	I have an orange BUTTON/PENCIL now
	I have a yellow BOTTLE/TABLE now

These stimuli were presented as coloured pictures on laminated cards. In the first condition, sentence-medial adjectives were focused (hereafter referred to as ADJ-FOC). In the second condition, sentence-final nouns were focused (hereafter referred to as N-FOC). In the third condition sentence medial nouns were focused (hereafter referred to as N-FOC-now).

2.3. Procedure

Participants took part in an elicited production task. The task engaged the participants by inviting them to play a language card game (referred to as 'Snap') with a female AusE-speaking experimenter. Before the test session, the experimenter explained how the game was played and went through three practice trials to familiarize participants with the game procedure before the test session.

During the test session, the experimenter and participant each received a deck of 16 cards with coloured stimuli images on the front of each card and a yellow star on the back of some. All the cards were held so that the images were concealed from the other player. For each trial, the experimenter would reveal the top card of their deck and produce the name of the stimulus item in the carrier sentence with neutral intonation, for example, '*I have a green moon*'. The participant would then reveal the top card of their deck and produce their item in the same carrier sentence. In each case the participant's item would differ from the experimenter's item in either type or colour of object, but not both. The participant was encouraged to emphasize the attribute/property which differed between the two pictures. Thus the participant was expected to produce adjective focus (e.g. '*I have a GREY ball*') on the trials where their item differed from the experimenter's in colour, and noun focus (e.g. 'I have a green SHOE') when their item differed from the experimenter's in object type. After that, both the experimenter and the participant put their cards face-up on the table and counted to three before turning their cards over. When one of the cards had a star marked on the back, whichever player called 'Snap' first received one point. This continued until both players used up all their cards.

The set of 16 stimuli were presented twice in different orders. For the first repetition the carrier phrase 'I have a/an X' was used. The second repetition used the carrier phrase 'I have a/an X now'. The adjective-focused sentences in the second repetition were treated as fillers, as sentence-medial adjective focus had already been elicited in the first repetition, giving a total of 24 sentences for analysis (see Table 1). The presentation order of cards was counterbalanced for all participants such that half the participants were presented the pictures in one order, and the other half in the reverse order. The responses were audio-recorded onto a PC using Audacity (audio recording software) at a sampling rate of 44.1 kHz, with a Behringer C2 condenser microphone.

2.4. Acoustic coding

Productions of the adjective + noun phrases were annotated and segmented in Praat [12], coding for the onset and offset of both the adjectives and the nouns. The onset consonants consisted of five types: (a) a stop/plosive, (b) a palatal glide, (c) a nasal (d) a fricative, and (e) no onset consonant. The coda consonants consisted of the following four types: (a) an affricate, (b) a lateral, (c) a nasal, and (d) no coda consonant. The coding criteria were therefore based on the ease of identifying the beginning and the end of the adjective and noun stimuli. When the onset consonant was a plosive/stop consonant, the beginning of the onset was indicated by the onset of the burst release. In items containing an onset glide, we used a pause (if present) and voicing to identify the onset. An additional cue was to use F2 transition to the palatal glide from a schwa in the preceding word. When the item contained a nasal consonant, onset of the nasal resonance was used as the cue. The beginning of high energy noise was used to identify items containing a fricative onset consonant. When there was no onset consonant, we used the onset of clear F2 and voicing to mark the beginning of the word.

To identify the end of words, we used the offset of the fricative portion for affricates. The end of the lateral coda consonant was based on the offset of voicing and F2 with minimal energy. The offset of nasal resonance was used to identify the end of words containing a nasal coda. Voicing and F2 offset were used as cues to the end of words containing no coda consonants.

3. Results

3.1. Adult duration

Word durations of the target adjectives and nouns in focus vs. non-focus positions were extracted. A repeated measures ANOVA was conducted, with ‘Focus’, ‘Number of syllables’ and ‘Word category’ as factors. Alpha was set at .05.

There was a significant main effect of ‘Focus’ (F = 52.148, df = 1, 7, p < .0001), and a significant main effect of ‘Number of syllables’ (F = 53.913, df = 1, 7, p <.0001). There was also a significant 3-way interaction among ‘Focus’, ‘Number of syllables’ and ‘Word category’ (F = 6.818, df = 1, 7, p =.035).

Not surprisingly, disyllabic target words were generally longer than monosyllabic words (344ms vs. 286ms). As predicted, focused word duration was longer than its non-focused counterpart (343ms vs. 287ms) (see Figure 1). The 3-way interaction resulted from larger accentual lengthening for monosyllabic nouns than monosyllabic adjectives (74ms vs. 44ms). However, the pattern was reversed for disyllabic focused words, with less accentual lengthening on disyllabic nouns than disyllabic adjectives (31ms vs. 73ms). Since monosyllabic nouns also occurred utterance-finally, this suggests that phrase-final lengthening might interact with focus-related lengthening on monosyllabic nouns.

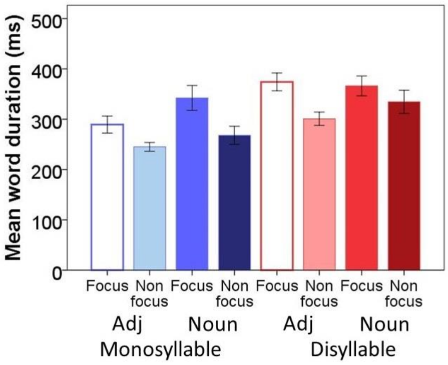


Figure 1. *Mean duration (ms) of focused vs. non-focused adjectives and nouns containing either monosyllables or disyllables in the adult group, with +/- 1 SE.*

To tease apart these two sources of lengthening, we performed another repeated measures ANOVA on focused nouns in utterance-final vs. utterance-medial positions. Two independent variables were included in the analysis: ‘Position’ and ‘Number of syllables’. Alpha was set at 0.05. If focus-related lengthening is further modulated by utterance-final position, we expected the focused nouns to be longer in final positions than non-final positions.

There was a significant main effect of ‘Number of syllables’ (F = 15.015, df = 1, 7, p <.006). However, counter to our prediction, no significant main effect of ‘Position’ was found. Yet there was a significant 2-way interaction between ‘Number of syllables’ and ‘Position’ (F = 7.592, df = 1, 7, p =.028). The interaction arose because the focused noun was longer in final than non-final position only when the nouns were monosyllabic (342ms vs. 284ms). However, when the nouns were disyllabic, there was no durational difference between final and non-final position (366 ms vs. 348 ms) (see Figure 2).

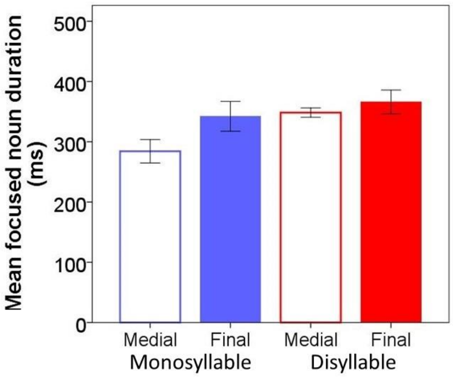


Figure 2. *Mean duration (ms) of focused noun in sentence-medial vs. sentence-final positions in the adult group, with +/- 1 SE.*

3.2. Child duration

Using word durations of focused adjectives and nouns as the dependent variable, we again conducted a repeated measures ANOVA to examine accentual lengthening on target adjectives and nouns in the children’s data, with the same three factors: ‘Focus’, ‘Number of syllables’ and ‘Word category’. Alpha was set at 0.05.

Unlike in adults, there was no significant main effect of ‘Focus’ (F = .007, df = 1, 7, p =.937). However, similar to the adults, there was a significant main effect of ‘Number of syllables’ (F = 59.402, df = 1, 7, p <.0001). Like the adults, there was also a significant 2-way interaction between ‘Focus’ and ‘Number of syllables’ (F = 11.984, df = 1, 7, p =.011), and a significant 3-way interaction among ‘Focus’, ‘Number of syllables’ and ‘Word category’ (F = 6.623, df = 1, 7, p =.037).

Unlike the adults, children did not show robust accentual lengthening. As expected, they did show longer duration for disyllabic words than monosyllabic words (410 ms vs. 315 ms). In contrast to adults, children showed an interaction between ‘Focus’ and ‘Number of syllables’. This interaction arose because accentual lengthening only took place for monosyllabic, not disyllabic target words.

The significant 3-way interaction arose because all target words underwent accentual lengthening except the focused disyllabic adjectives. Instead of lengthening, the focused disyllabic adjectives were shortened relative to their non-focused counterparts. (see Figure 3)

Since the focused nouns occurred utterance-finally, the observed accentual lengthening could be utterance-final lengthening in disguise. Therefore, a separate repeated measures ANOVA was performed to investigate children’s ability to use the durational cue to signal focus. Two independent variables were included in the analysis: ‘Position’ and ‘Number of syllables’. Alpha was set at 0.05.

There was a significant main effect of ‘Number of syllables’ (F = 79.724, df = 1, 7, p <.0001). Counter to the adult patterns and our prediction, there were neither a main effect of ‘Position’ nor a ‘Position-Number of syllables’ interaction (see Figure 4).

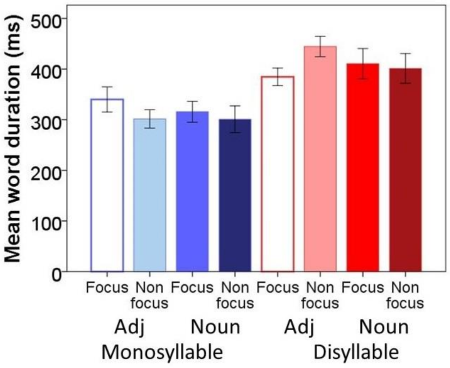


Figure 3. *Mean duration (ms) of focused vs. non-focused adjectives and nouns containing either monosyllables or disyllables in the child group, with +/- 1 SE.*

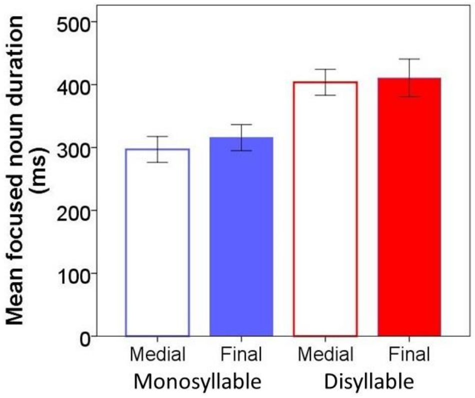


Figure 4. *Mean duration (ms) of focused noun in sentence-medial vs. sentence-final positions in the child group, with +/- 1 SE.*

4. Discussion

The findings showed that children can employ accentual lengthening to signal focus; however, this ability is not the same as or as robust as that observed in the adults. As predicted in H_1, adults exhibited accentual lengthening in both monosyllabic and disyllabic words. In partial support of H_2, children, however, implemented accentual lengthening only on monosyllabic, not disyllabic words. In addition, the adults showed accentual lengthening on focused disyllabic adjectives; whereas the children shortened them. This indicates that children are still inconsistent in using the durational cue to signal focus. In other words, children are still learning how to generalize accentual lengthening across the number of syllables, and its subtle acoustic realization.

It is interesting that the interaction of accentual lengthening and phrase-final lengthening is contingent on the number of syllables within a word. The adults showed additional phrase-final lengthening effects on focused monosyllabic words, but not focused disyllabic words. This partially supports H_3. It is possible that phrase-final lengthening might be present on the second syllable of the disyllabic word, which requires further analysis. Unlike the adults, *no* additional phrase-final lengthening was observed in the children. When children implement accentual lengthening, phrase-final lengthening seems to disappear. This suggests that children are still learning how to weigh duration to signal focus and sentence position separately in an adult-like manner.

5. Conclusion

This study showed that children aged between 5 and 6 years can use accentual lengthening, however, this ability is not as robust as what previous studies have suggested, and far from being adult-like. Perhaps the disparity between the production and perception of focus in children is not as anomalous as previously thought, given that children are still learning how to use the appropriate acoustic cues to signal focus in production.

6. Acknowledgements

This work was supported by the ARC Laureate Fellowship [FL130100014] awarded to the last author. We would like to thank our participants for taking part in the research, Amy German and Cecily Mccowage for their assistance in data collection and annotation, members of the Child Language Lab and the two anonymous reviewers for their helpful comments

7. References

[1] Cooper, W. E., Eady, S. J. and Mueller, P. R. “Acoustical aspects of contrastive stress in question-answer contexts”, Journal of Acoustical Society of America, 77:2142-2156, 1985

[2] Fry, D. B. “Duration and intensity as physical correlates of linguistic stress”, Journal of the Acoustical Society of America, 27: 765-768, 1955.

[3] Kochanski, G., Grabe, E., Coleman, J., and Rosner, B. “Loudness predicts prominence: fundamental frequency lends little”, Journal of the Acoustical Society of America, 118:1038-1054, 2005.

[4] Eady, S. J. and Cooper, W. E. “Speech intonation and focus location in matched statements and questions”, Journal of the Acoustical Society of America, 80: 402-416, 1986.

[5] Speer, S. R. and Ito, K. “Prosody in first language acquisition – acquiring intonation as a tool to organize information in conversation”, Language and Linguistics Compass, 3:90-110, 2009.

[6] Turk, A. E. and White, L. “Structural influences on accentual lengthening in English”, Journal of Phonetics, 27:171-206, 1999.

[7] Turk, A. E. and Shattuck-Hufnagel, S. “Multiple targets of phrase-final lengthening in American English words,” Journal of Phonetics, 35:445-472, 2007.

[8] Hornby, P. A. and Hass, W. A., “Use of contrastive stress by preschool children”, Journal of Speech and Hearing Research, 13: 395-399, 1970.

[9] Snow, D. “Phrase-final syllable lengthening and intonation in early child speech”, Journal of Speech and Hearing Research, 37:831-840, 1994.

[10] Wells, B., Peppé, S. and Goulandris, N. “Intonation development from five to thirteen”, Journal of Child Language, 31: 749-778, 2004.

[11] Patel, R. and Brayton, J. T. “Identifying prosodic contrasts in utterances produced by 4-, 7- and 11-year old children”, Journal of Speech, Language and Hearing Research, 52:790-801, 2009.

[12] Boersma, P. and Weenink, D. Praat: doing phonetics by computer. http://www.praat.org/

A Comparative Ultrasound Study of Coronal Consonants in Arrernte and Kannada: Manner Contrasts

Marija Tabain[1], *Alexei Kochetov*[2], *Richard Beare*[3,4] *& N. Sreedevi*[5]

[1]La Trobe University, Australia
[2]University of Toronto, Canada
[3]Monash University, Australia
[4]Murdoch Children's Research Institute, Australia
[5]All India Institute of Speech and Hearing, India

m.tabain@latrobe.edu.au, al.kochetov@utoronto.ca, richard.beare@monash.edu, nsreedevi@aiishmysore.in

Abstract

We present ultrasound data from four speakers of the Australian language Arrernte, and ten speakers of the Dravidian language Kannada. Our focus is on the coronal consonants of these languages, and in particular on the manner contrasts (stop, nasal and lateral) for the various places of articulation. The places of articulation are dental, alveolar, retroflex and palatal for Arrernte; and dental/alveolar and retroflex for Kannada. We show that the tongue back is consistently more back for the lateral manner of articulation, and almost always more forward for the nasal manner. We discuss the possible reasons for these results.

Index Terms: ultrasound, coronal consonants, manner contrasts, Australian languages, Dravidian languages

1. Introduction

The languages of Aboriginal Australia and the Dravidian languages of India are remarkably similar in terms of phonemic structure, yet very little work has compared these two groups of languages directly. In this study we present a comparison of the relative tongue back positions for the coronal consonants of two languages, Arrernte (Australian) and Kannada (Dravidian), with a particular focus on the manner contrasts – stop, nasal and lateral – at each place of articulation. Arrernte is a language of Central Australia, spoken by about 2000 people in and around the administrative township of Alice Springs (*Mparntwe*). It has four coronal places of articulation: dental, alveolar, retroflex and alveo-palatal. The dental and palatal sounds are classified as laminal consonants, and the alveolar and retroflex sounds are classified as apical consonants [1,2].

Kannada is a Dravidian language spoken mainly in the South Indian state of Karnataka by over 35 million people [3]. The main difference within Kannada coronals is between retroflexes /ʈ ɖ ɳ ɭ/ and non-retroflexes, with non-retroflexes being either laminal dental (for stops /t̪ d̪/) or apico-laminal alveolar (for nasals and laterals /n l/) [4,5]. The set also includes alveolopalatal affricates (/ʧ ʤ/, not examined here). All these consonants occur as singletons and geminates.

In this paper, we consider ultrasound data for the stop, nasal and lateral coronal consonants in these languages. Preliminary examination of our data suggested that the differences resided primarily in the posterior half of the tongue as imaged by ultrasound, and this is therefore the portion of the tongue that we focus on in this study. However, since the Kannada non-retroflex consonants have different places of articulation according to manner – dental for the stops, and alveolar for the nasals and laterals – we do not directly compare the Kannada non-retroflex stop with the Kannada non-retroflex lateral and nasal.

2. Method

2.1. Speakers and Recordings

2.1.1. Arrernte

Seven female speakers of Arrernte were recorded to ultrasound using the Telemed Echo Blaster 128 CEXT-1Z, the Articulate Instruments stabilization helmet [6, 7], the Articulate Instruments pulse-stretch unit, and the AAA software version 2.16.07 [8]. In addition we used an MBox2 Mini soundcard, a Sony lapel microphone (electret condenser ECM-44B), and an Articulate Instruments Medical Isolation Transformer. The ultrasound machine, sync pulse, sound card and a software dongle were connected via USB to a Dell Latitude E6420 laptop running Windows software. Typical frame rate was 87 f.p.s., using a 5-8 MHz convex probe set to 7 MHz, a depth of 70 mm and a field of view of 107.7 degrees (70%). An eighth potential speaker was not recorded because we were unable to see a clear outline of her tongue.

Recordings took place either in a hotel room in Alice Springs (five speakers) or in the staff-room of the Santa Teresa school, 85 km south-east of Alice Springs (two speakers).

Speakers read a list of 92 Arrernte words designed to present the four coronal places of articulation for the oral stop, nasal and lateral consonant series. Some of these words illustrated homorganic nasal+stop, stop+nasals or lateral+stop clusters. Wherever possible, surrounding vowels were the central vowels /a/ or /ə/, which are the most common vowels in Arrernte (Arrernte has three phonemic vowels, /a ə i/, and a fourth vowel [u] which occurs as a result of rounding on a consonant – rounded consonants were avoided in the list, though were not entirely absent). Where possible, the target consonants were illustrated both in stressed and in unstressed word position (note that schwa can be stressed in Arrernte).

The words were displayed on the laptop screen. Speakers were asked to say each word three times, or as often as possible within the 5-second recording window set by the Articulate Assistant software. Some speakers were able to produce four or five repetitions in each 5-second window. Each speaker read the list through at least once, and four speakers read through the list a second time. Some speakers chose not to produce a particular taboo word which was accidentally included on the list, and some speakers weren't sure of some words. Note also that the ultrasound machine does not begin recording until about 150 ms into the 5-second audio recording window – as a result, some repetitions were discarded because they were cut off by these limitations.

In the present study, we are only presenting data from four speakers – these particular speakers were chosen for these initial analyses because their tongue images were relatively clear for a large set of the target consonants. Note that two speakers presented in the current study were not literate in Arrernte, and were prompted by another (literate) speaker who was present in the room, and/or by the author MT providing an English gloss of the target word. One of these speakers read through the list once, and the other speaker read through the list twice.

A total of 3653 tokens were analyzed for this study. Three of the speakers had about 1000 tokens, while the fourth speaker had about 600 tokens.

2.1.2. *Kannada*

Ten Kannada speakers (5 females and 5 males, 21-26 years old) from Karnataka, India were recorded using a PI 7.5 MHz SeeMore probe by Interson Corporation (http://www.interson.com/) with a 90 degree field view and a depth of 10 cm at the frame rate of 15 f.p.s. The ultrasound image was displayed using the SeeMore software (version 1.3.02) on a Lenovo ThinkPad Edge E220 s laptop and synchronized with the audio (using a Sony DVDirect MC6 multi-function DVD recorder via a PC-to-TV converter) at 29.97 frames per second (the NTSC standard rate). The audio signal was captured at 48 kHz using an AT831b lavalier microphone and a Sound Devices USBPre2 pre-amp. The ultrasound probe was stabilized using Articulate Instruments stabilization helmet [6, 7].

The data were collected in a quiet room in the Speech Sciences Department at the All India Institute of Speech and Hearing, Mysore.

The materials consisted of Kannada VCːV words (where Cː is geminate) with consonants of various places and manners of articulation. The current study examines manner differences in dentals/alveolars and retroflexes in the [a_a] context, involving 6 meaningful words [at̪ːa] 'that side', [anːa] 'cooked rice', [alːa] 'not', [aʈːa] 'garret', [haɭːa] 'small stream', and [aɳːa] 'elder brother' (henceforth referred to as atta, anna, alla, aTTa, aLLa, and aNNa). The first vowel was typically pronounced as reduced in duration and raised to [ʌ]; the second vowel was a low central [a].

The words were written in the Kannada orthography and presented on a laptop screen 10 times each, with an inter-stimuli interval of 1 second. Words with geminates were selected to ensure that consonant closures were adequately captured by the ultrasound system despite the relatively slow frame rate.

A total of 591 tokens of the 6 target words (or on average 59 per speaker) were recorded, with 9 tokens skipped by some of the speakers.

2.2. Data processing

2.2.1. *Arrernte*

Acoustic data were labelled using the EMU speech software package [9] version 2.3, and tongue contours were tracked semi-automatically using the AAA software. Tracked tongue data were exported as text files with 42 x and y coordinates plus confidence levels (note that all confidence levels were accepted for the present study, given that much manual correction was involved in the tracking). Data were subsequently converted to Simple Signal File Format (SSFF) for compatibility with EMU. Analyses were conducted using EMU/R version 4.4, interfaced with the R statistical package version 3.1.2 [10].

2.2.2. *Kannada*

Individual frames were extracted from videos and the frames corresponding to consonant constrictions were identified by visual inspection with reference to the acoustic signal. Tongue contours were traced semi-automatically using EdgeTrak [11] and exported as text files with 100 x and y coordinates (with no confidence intervals).

2.2.3. *Analyses*

Tongue spline data were sampled at the acoustic offset of the consonant for Arrernte. For Kannada, tongue spline data were sampled at the frame with maximum consonant constriction – this tended to be the last frame of the acoustic consonant duration.

For both Arrernte and Kannada, smoothing spline ANOVAs (SS-ANOVAs) were calculated using the *gss* package in R [12]. Figures were created using the *ggplot2* package [13].

It will be recalled that all of the Kannada stimuli were inter-vocalic, whereas the Arrernte stimuli were mostly intervocalic, but also contained word-initial and -final tokens, as well as homorganic clusters. However, we chose to include all of the vocalic and prosodic contexts for the Arrernte analysis, since including only the intervocalic context for Arrernte would have left us with no data for the palatal nasal (the sequence /aˈɲə/ is a taboo word in Arrernte). We did examine the intervocalic-only data for Arrernte, and can confirm that the patterns presented below are quite similar for intervocalic context and for all contexts.

In addition, it should be noted that no attempt was made to separate the apicals (alveolar or retroflex) according to prosodic context in the Arrernte data: the apical contrast is in principle neutralized in word-initial position, and the retroflex is more prototypically retroflex when it is in unstressed position [14, 15].

Based on [16], we examined the posterior half of the tongue spline. The splines were considered significantly different if there was no overlap for over 2/3 of this portion – labelled '>' in the table below. A trend was defined if there was no overlap for over 1/2 of this portion of the tongue – labelled '≥' in the table below). SS-ANOVA figures were examined by both the first and the second author of this paper.

3. Results

Figure 1 gives an example of tongue splines for one speaker of Arrernte, and Figure 2 gives an example of tongue splines for one speaker of Kannada. It can be seen that the rear-most 50% of the tongue for the lateral (red line) is more posterior than for the stop (blue line) and the nasal (green line) in all of these plots. In addition, for this Kannada speaker's data, it can be seen that the rear-most 50% of the tongue is more anterior for the nasal (green line) than for the stop or for the lateral, for both the alveolar and the retroflex places of articulation. However, this is not the case for the Arrernte speaker – it can be seen that the nasal is more forward only for the dental data for this speaker, and not for the other three places of articulation.

It should also be remarked that the dental stop for the Kannada speaker in Figure 2 (and also for the other Kannada speakers not shown here) has a much flatter tongue body than the alveolar nasal and lateral – recall that the dental stop and the alveolar lateral and nasal are classified as non-retroflexes. This flatter tongue body for the Kannada dental is consistent with the flatter tongue body for the stop, nasal and lateral dentals for the Arrernte speakers (the three other speakers are not shown here). This flatter tongue body is often accompanied by a more posterior tongue back position: this is why, as we noted in the Introduction, we do not directly compare the Kannada non-retroflex stop with the Kannada non-retroflex lateral and nasal.

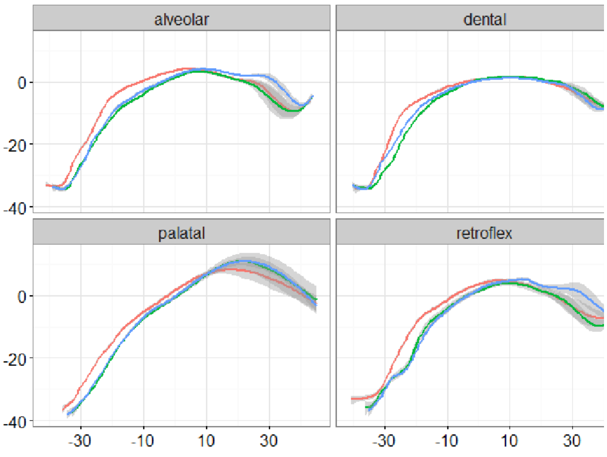
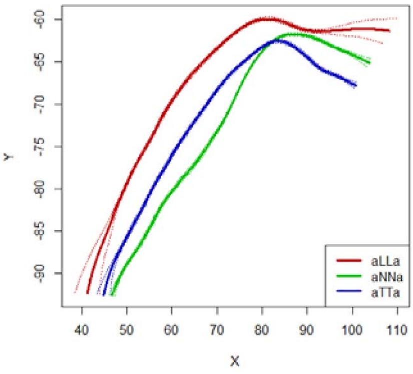


Figure 1: *Tongue contours for Arrernte speaker AF2, sampled at the offset of the consonant. The red line denotes a lateral consonant, the green line denotes a nasal, and the blue line denotes a stop. The grey shadows surrounding each contour represent the confidence intervals generated by the SSANOVA.*



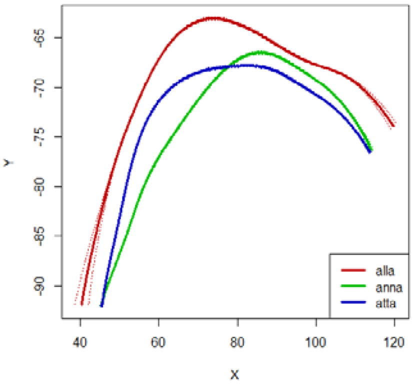


Figure 2: *Tongue contours for Kannada speaker KF2, sampled at the offset of the consonant. Top panel: dental/alveolars. Bottom panel: retroflexes. The red line denotes a lateral consonant, the green line denotes a nasal, and the blue line denotes a stop. The dotted lines surrounding each contour represent the confidence intervals generated by the SSANOVA.*

Table 1. *Classification of tongue posterior spline location for four speakers of Arrernte and ten speakers of Kannada. '>' denotes that the tongue position is more back through at least 2/3 of the rear-most 50% of the spline. '≥' denotes that the tongue position is more back through at least half (but less than 2/3) of the last 50% of the spline. '=' denotes no difference in tongue position for at least half of the rear-most 50% of the spline. 'L' denotes lateral, 'S' denotes stop, and 'N' denotes nasal. Cells in red denote a pattern that does not fit in with the dominant pattern.*

Arrernte	**Dental**	**Alveolar**	**Retroflex**	**Palatal**
AF1	L > S = N	L > S = N	L > S = N	L > S = N
AF2	L > S > N	L > S = N	L > S = N	L > S = N
AF3	L = S > N	L > S = N	L > S = N	L = S = N
AF4	L > S = N	L > S = N	L > S = N	L > S > N
Kannada		**Alveolar**	**Retroflex**	
KF1		L > N	S > L > N	
KF2		L > N	S > L > N	
KF3		L > N	N > L = S	
KF4		L > N	L > S > N	
KF5		L > N	S = N ≥ L	
KM1		L > N	L > S > N	
KM2		L > N	L > S > N	
KM3		L > N	L = S > N	
KM4		L > N	L > S = N	
KM5		L > N	L = S = N	

Table 1 provides a list of the relative placement of the tongue back for the stop, nasal and lateral positions for each place of articulation, for each speaker. It can be seen that for Arrernte, the lateral manner is more posterior than the nasal manner (L > N) for 14 out of the 16 comparisons, and more posterior than the stop manner (L > S) for 15 out of the 16 comparisons. By contrast, the stop manner is more posterior than the nasal manner for only three of the 16 comparisons (S

> N) – in all other cases, there is no difference between the stop and the nasal (S = N).

For the Kannada comparisons, it can be seen that the lateral is posterior to the nasal for all ten of the alveolar comparisons (L > N). The same is true for seven out of the ten retroflex comparisons (the three exceptions are marked in red, with one exception showing no difference between lateral and nasal, and the other two exceptions showing the reverse pattern). By contrast, the relative placement of the stops in this tongue back placement hierarchy varies greatly from speaker to speaker – nevertheless, it can be remarked that the stop is posterior to the nasal (S > N) in six out of ten cases. However, lateral is posterior to stop (L > S) for only four out of ten cases, with three cases showing no difference between the two (L = S). In three cases, stop is posterior to lateral (S > L).

4. Conclusion

There is a strong preference for the tongue back to be further forward for the nasal than for the lateral consonants, for all places of articulation. This may be due to the need to avoid contact between the tongue back and the velum, which has been lowered in order to allow airflow through the nasal cavity. As pointed out by Fant [17], it is possible for the back of the tongue to contact the centre of the uvula during nasal consonant production, with air flowing along the sides of this connecting point — this may occur, for example, with a high tongue position and a fully lowered velum. Such a configuration may set up a different set of oral and nasal resonances, which may not be desirable in the case of a place-rich consonant system [18]. By contrast, the hydrostatic nature of the tongue may mean that as the tongue sides are lowered for lateral production, the back of the tongue is pushed further back to compensate for this – or perhaps, lateral production involves active elongation of the tongue by dorsal retraction [19].

What is not clear in our study is the relative placement of the stop manner of articulation in the hierarchy of tongue back placements. Whilst in Arrernte there is a tendency for the stops to pattern with the nasals, this is not quite so clearly the case for Kannada. Whether these differences are genuine differences in the languages, or whether they are an artefact of the different methodologies that we used for data collection in the two languages, is a question for future studies.

5. Acknowledgements

We would like to thank our Arrernte and Kannada speakers for their interest in language work; Gavan Breen and Midula Kasim for assistance with participant recruitment and data collection, and Casey Tait for (Arrernte) data annotation. This research was supported by an Australian Research Council Future Fellowship to Marija Tabain, an All-India Institute of Speech Hearing (AIISH) Research Fund to Dr. N. Sreedevi and Alexei Kochetov, and a Social Sciences and Humanities Research Council of Canada Standard grant to Alexei Kochetov. Ethical approval for this research was granted at La Trobe University, at AIISH, and at the University of Toronto.

6. References

[1] G. Breen and V. Dobson. "Central Arrernte," *Journal of the International Phonetic Association,* vol. 35, pages 249-254, 2005.

[2] J. Henderson, *Topics in Eastern and Central Arrernte grammar*. Lincom Europa: Germany, 2013.

[3] M. Lewis. (ed.). *Ethnologue: Languages of the world*, 16th edition. Dallas, TX: SIL International. Online version: http://www.ethnologue.com/. 2009.

[4] H. Schiffman, H. *A reference grammar of spoken Kannada*. Seattle: University of Washington Press. 1983.

[5] U. P. Upadhyaya, *Kannada phonetic reader*. Central Institute of Indian Languages, Mysore. 1972.

[6] Articulate Instruments Ltd., *Ultrasound Stabilisation Headset Users' Manual: Revision 1.4*. Edinburgh, UK: Articulate Instruments Ltd., 2008.

[7] J. Scobbie, A. Wrench, A., and M. van der Linden. "Head-probe stabilization in ultrasound tongue imaging using a headset to permit natural head movement," *Proceedings of the 8th International Seminar on Speech Production*, pp. 373-376. 2008.

[8] Articulate Instruments Ltd., *Articulate Assistant Advanced User Guide: Version 2.14*. Edinburgh, UK: Articulate Instruments Ltd., 2012.

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

[10] 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.

[11] M. Li, C. Kambhamettu and M. Stone. "Automatic contour tracking in ultrasound images," *Clinical Linguistics and Phonetics, 19*, 545–554. 2005.

[12] C. Gu, "Smoothing Spline ANOVA Models: R Package gss," *Journal of Statistical Software*, vol. 58(5), pages 1-25. URL http://www.jstatsoft.org/v58/i05/, 2014.

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

[14] M**.** Tabain and R. Beare. "An EPG and EMA study of apicals in stressed and unstressed position in Arrernte," *18th International Congress of the Phonetic Sciences.* Glasgow: Scotland, 2015.

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

[16] A. Kochetov, N. Sreedevi, M. Kasim and R. Manjula. "Spatial and dynamic aspects of retroflex production: an ultrasound and EMA study of Kannada geminate stops," *Journal of Phonetic, 46,* 168-184. 2014.

[17] G. Fant, *Acoustic Theory of Speech Production,* 2nd ed. (Mouton, The Hague), 1970.

[18] M. Tabain, A. Butcher, G. Breen and R. Beare, "An acoustic study of nasal consonants in three Central Australian languages," *Journal of the Acoustical Society of America, 139,* 890-903, 2016.

[19] P. Ladefoged & I. Maddieson, *The Sounds of the World's Languages.* (Blackwell: Oxford, UK ; Cambridge, Mass.), 1996.

Verbal Reduplication and Minimal Words in Kaytetye

Forrest Panther[1], Mark Harvey[1], Harold Koch[2], Myfany Turpin[3], Michael Proctor[4]

[1]University of Newcastle, Australia
[2]Australia National University, Australia
[3]University of Sydney, Australia
[4]Macquarie University, Australia

ForrestAndrew.Panther@uon.edu.au, Mark.Harvey@newcastle.edu.au, Harold.Koch@anu.edu.au, Myfany.Turpin@sydney.edu.au, Michael.Proctor@mq.edu.au

Abstract

This paper examines one type of reduplication in Kaytetye, a Pama-Nyungan language of central Australia, which raises important questions for phonological theory. Kaytetye has a reduplication construction, which expresses associated motion, where the final VCV of verb roots is reduplicated to form a second prosodic word: e.g. *alarre* "hit", *al**arre**-lp-**arre*** "hit on the way". This reduplication construction provides key evidence for the shape of the minimal word in Kaytetye, and for current analyses of reduplication.

Index Terms: Reduplication, Australian Phonology, Arandic

1. Introduction

A distinction has been proposed between morphological reduplication and syntactic reduplication [1-3]. This distinction is primarily based on the morpho-syntactic category of the overall base + reduplicant construction: word level vs. phrase level or higher. Morphological reduplication is word level, whereas syntactic reduplication is phrase level or higher. That is, in morphological reduplication, the reduplicant compounds or affixes with the base, whereas it does not in syntactic reduplication. Several authors also argue for a second criterion, that while morphological reduplication must produce a well-formed phonological unit, syntactic reduplication must produce a well-formed grammatical unit [1-3]. The proposed distinction between these two types of reduplication is not supported in Morphological Doubling Theory, which is explored in section 4.1 [4, 5]. We examine here a type of reduplication pattern that addresses this distinction, found in Kaytetye, an Arandic (Pama-Nyungan) language of Central Australia [6, 7].

Most of the literature on reduplication examines morphological reduplication, which involves affixing a partial or total reduplicant to the base [2, 8-11]. The literature on syntactic reduplication is more limited, but it is discussed by several authors [1, 15-16]. (1) is an example of syntactic reduplication, showing reduplication of the verb "come" in Lango [4]:

1) àbíno àbínə àwó'ró
1SG.come.PERF come.GER yesterday
"I did come yesterday" [4]

This construction in (1) is analysed as syntactic reduplication because the reduplicant is delimited by syntactic and not phonological criteria. The reduplicant is a verb, and the tense/aspect inflections of the base and reduplicant are distinct. These are considered to have separate lexical inputs, and represent two different words. Reduplication at this level is considered to be of whole, grammatical units (Fongbe is a possible exception to this, with optional truncation of the reduplicant in some speakers. See [4, 5]). There is, as of yet, no description of obligatory partial reduplication at a syntactic level.

The Kaytetye reduplication construction we describe here patterns in this way: it is reduplication at the phrasal level. We show this pattern to be strictly partial reduplication, and demonstrate that the partial reduplicant is a minimal word (see [14] for a similar pattern in Eastern and Central Arrernte). We examine the structure of the minimal word in Kaytetye in Section 2. In Section 3, we examine the reduplication pattern in detail. Section 4 will discuss this pattern in light of current models of reduplication and prosodic structure.

2. Minimal Words in Kaytetye

The *minimal word* is a restriction on the minimum phonological structure required for words in a given language [8, 18, 19]. Phonological theory posits that minimal words must be at least bimoraic, and this can be satisfied by either a heavy monosyllable or a disyllabic foot [9, 17]. There is variation in Australian languages on what kind of minimal word they permit, but many Pama-Nyungan languages require a disyllabic minimum [20-22]. This disyllabic minimum – CV(C)CV(C) – is described as the "prototypical phonological word" of Australian languages [21]. The edges of this prototypical word are marked, with several phonological contrasts neutralized and a lack of clustering at the initial and final positions [20, 21].

Kaytetye conforms to the general Pama-Nyungan pattern in that no prosodic word is smaller than disyllabic, and almost all lexical roots are minimally disyllabic (There are very few exceptions). However, approximately 75% of Kaytetye words begin with an unstressed vowel (often /ɐ/) [23]. All Kaytetye words also end in a non-contrastive vowel which shows considerable variation in realisation [24]. This raises questions about the utility of the consonant-initial prototypical word template CV(C)CV(C) for Kaytetye. Given the tendency for vowels to occupy word edges, the Kaytetye minimum word may be better analysed as VC(C)V (hereafter summarized as VCV for convenience).

There is evidence from affixal allomorphy that the minimal word in Kaytetye is VCV. The Ergative suffix distinguishes VCV roots from all other roots, including CVCV roots. With VCV roots, the Ergative is *-nge* (2), whereas with other roots, including CVCV roots, it is *–le* (3) [7: p.60]:

2) VCV roots:
arrke-nge
/aɾkə-ŋə/
Sun–ERG
“the sun”

3) CVCV and longer roots:
relhe-le
/ɹəl̪ə-lə/
Woman-ERG
“the woman”

The reduplicant in the associated motion construction also has a VCV shape. We propose that it has this shape because the reduplicant is constrained to conform to a minimum word.

3. Phrase Level Reduplication in Kaytetye

Phrase-level constructions in Kaytetye differ from word-level constructions, such as compounds. A key test for distinguishing phrase-level from word-level constructions is the placement of clitics. In word-level constructions, such as compounds, clitics cannot be placed in a medial position (= indicates a clitic boundary):

4) akelperr-elperre=rtame
/ɐkəlpəɾ-əlpəɾə=ʈɐmə/
head-flat=EMPH
“a big head”

*ake=rtame=elperre (unattested)

However, phrasal constructions do allow medial clitics, as illustrated in (5):

5) weye=lk=aherre
/wijə=lk=ɐɰəɾə/
meat=then=kangaroo
“then kangaroo meat”

The reduplication construction patterns in the same way as phrasal constructions. Examples (6) and (7) illustrate the reduplication construction with medial clitic placement.

6) ile-lpe=lk=ile-nye
/ilə-lpə=lk=ilə-ɲə/
get-LIG=then=RED-PST
“then (someone) got it on the way”

7) atnywe-lpe=lk=atnywe-ye
/ɐˤɲu-lpə=lk=ɐˤɲu-jə/
enter-LIG=then=RED-FUT
“Then (it) will go down (into)”

As previously discussed, proposed models of syntactic reduplication predict the reduplication of syntactic units. This would mean the reduplication of whole grammatical words, as in (8), or the total reduplication of roots, as in (9):

8) *ile-lp-ilelpe-nye
/ilə-lp-iləlpə-ɲə/
get-LIG-RED-PST
“then (someone) got it on the way”

9) *alarre-lp-alarre-nye
/ɐlɐɾə-lp-ɐlɐɾə-ɲə/
hit-LIG-RED-PST
“hit on the way”

These predicted forms are not attested. Rather, the attested forms show reduplication only of the final VCV of the root.

VCV Roots
10) alwe-nke
/ɐlu-nkə/
Chase-PRES
“chase(s)”

alwe-lp-alwe-nke
/ɐlu-lp-ɐlu-nkə/
Chase-LIG-RED-PRES
“catch(es) up with something”

CVCV Roots
11) kwathe-nke
/kwɐt̪ə-nkə/
drink-PRES
“drink(s)”

kwathe-lp-athe-nke
/kwɐt̪ə-lp-ɐt̪ə-nkə/
drink-LIG-RED-PRES
“drink(s) on the way”

Longer Roots
12) alarre-nke
/ɐlɐɾə-nkə/
hit-PRES
“kill(s); hit(s)”

alarre-lp-arre-nke
/ɐlɐɾə-lp-ɐɾə-nkə/
hit-LIG-RED-PRES
“kill(s) (something) on the move”

It is important to note that CVCV roots do not show reduplication of the initial C, as shown in (11). Given the prototypical Pama-Nyungan word template CVC(C)V(C), the predicted reduplication for CVCV roots would be that in (13).

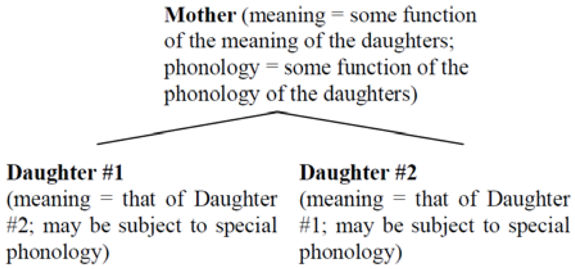
13) *kwathe-lpe-kwathe-nke
/kwɐt̪ə-lpə-kwɐt̪ə-nkə/
drink-LIG-RED-PRES
“drink(s) on the way

4. Discussion

The reduplication construction provides evidence bearing on two important questions: First, how to current models of reduplication account for this pattern? Second, what is the nature of the minimal word in Kaytetye, and what are the implications of this for prosodic theory more generally?

4.1 Analyses of reduplication

The Kaytetye reduplication construction is syntactic, but the reduplicant is phonologically conditioned. This offers a challenge to theories which distinguish morphological from syntactic reduplication. The Kaytetye construction also offers a challenge to theories which do not, such as Morphological Doubling Theory (MDT) [4, 5]. MDT proposes that reduplication is the doubling of a lexeme, and both the base and reduplicant have separate inputs, and are shaped by separate co-phonologies. Fig. 1 illustrates the relationship between base and reduplicant under this model:



Mother (meaning = some function of the meaning of the daughters; phonology = some function of the phonology of the daughters)

Daughter #1 (meaning = that of Daughter #2; may be subject to special phonology)

Daughter #2 (meaning = that of Daughter #1; may be subject to special phonology)

Figure 1: *Illustration of reduplication under the Morphological Doubling Theory (from pg. 1, [4])*

Under an MDT analysis, the daughters of *kwathe-lp-athe-nke* 'drink-LIG-RED-PRES' are *kwathe-lpe* 'drink-LIG' [Daughter 1] and *kwathe-nke* 'drink-PRES' [Daughter 2]. The truncated output form for Daughter 2, *athe-nke*, must be specified by the co-phonology for Daughter 2. The truncation can be described as the right aligned truncation of the root to a minimal word. However, no phonology operating at the level of a complex word, such as *kwathe-nke*, can produce the right-aligned truncation of a left sub-constituent within that complex word. The right-aligned truncation can only be analyzed at the level of the root *kwathe*, which is not the level of Daughter 2 *kwathe-nke*. As such, MDT does not offer an analysis of the Kaytetye reduplication construction.

4.2 Minimal words and prosodic structure

We have shown in sections 2 and 3 that there are two independent patterns which identify the phonological shape VCV as a target in Kaytetye. The first is the allomorphy of the ergative suffix. The second is the reduplicant in the associated motion construction. The minimal word is commonly identified as a target in nominal allomorphy in Australian languages [19]. The minimal word is also commonly identified as a target in reduplication [8, 16]

The reduplicant attracts stress independently. In Kaytetye, the standard pattern for stress placement is that the first vowel preceded by a consonant is stressed, and thereafter every second vowel is stressed [23]. Given this standard pattern, the predicted stress placement for *antheyayte-lp-ayte-nke* 'climbs up on the way' is that in (14), with three stresses.

14) *$_{\textbf{Wd}}$[an$_{\textbf{Ft}}$[ˈtheyay]$_{\textbf{FtFt}}$[ˈte-lp-ay]$_{\textbf{FtFt}}$[ˈte-nke]$_{\textbf{Ft}}$]$_{\textbf{Wd}}$
climb-LIG-RED-PRES
"climbs up on the way"

However, the actual stress placement is that in (15), with only two stresses.

15) $_{\textbf{Wd}}$[an$_{\textbf{Ft}}$[ˈtheyay]$_{\textbf{Ft}}$te-lp]$_{\textbf{Wd}}$ $_{\textbf{Wd}}$[-ay$_{\textbf{Ft}}$[ˈte-nke]$_{\textbf{Ft}}$]$_{\textbf{Wd}}$
climb-LIG-RED-PRES
"climbs up on the way"

The sequence of two unfooted syllables in (15) is the standard outcome, if the construction is analyzed as consisting of two prosodic words.

The VCV shape satisfies the universal bimoraic requirement for a minimal word, and attracts stress. Therefore, we propose that the Kaytetye minimal word is VCV. This VCV minimal word shape is typologically interesting with its lack of word-edge consonants, particularly the onset. We note that this minimal word shape correlates with the preference for words of all lengths to be vowel-initial in Kaytetye (see [25, 26] for discussion on the diachronic loss of word-initial consonants). Of particular interest is the fact that CVCV does not pattern prosodically with VCV. Rather, CVCV patterns prosodically with VCVCV.

Current analyses of prosody, which posit that only rhyme constituents may be moraic, do not offer an obvious analysis for the Kaytetye data. Arrernte, a related language, shows very similar phenomena, and data from Arrernte has been an important factor in analyses of the syllable. Gordon posits that the syllable onset is moraic [27]. Breen and Pensalfini propose that syllabification is VC rather than CV: i.e. there are no underlying syllable onsets [28].

Gordon's general analysis is based in perception. In connected speech, a consonant enhances the perception of a following vowel. The degree of enhancement depends on the manner of articulation of the consonant, with voiceless consonants providing a stronger cue than voiced consonants. Vowels not preceded by a consonant are not enhanced. Languages vary as to whether and how the enhancement is phonologized: i.e. as to whether and how onset consonants are moraic and therefore contribute to the determination of prosodic structure.

Gordon's specific analysis of Arrernte is based in vowel length. Vowels in onsetless syllables are significantly shorter than average, and vowels in open syllables with onsets are significantly longer than average. Together, these two factors enhance the effectiveness of onset-based contributions to the perception of prominence. Further research is required to fully evaluate Gordon's analysis in relation to Kaytetye. Current data indicates that, prima facie, it could provide a plausible analysis of Kaytetye stress placement [24].

Gordon's analysis does not involve a radical departure from standard syllabic structure. By contrast, the VC analysis does involve a radical departure from standard syllabic structure. Further, data from Arrernte reduplication constructions plays a critical evidentiary role in supporting the VC hypothesis.

Although the VC analysis is a hypothesis about syllable structure, its most immediately evident consequences are in the analysis of word structure. The VC analysis posits that all words are vowel-initial and consonant-final. As discussed in section 2, 75% of Kaytetye words are vowel-initial, and all words may involve a final non-contrastive vowel. In phonetic realization, Kaytetye words may begin with [ɐ, i] or a consonant. The VC analysis proposes that all consonant-initial words have an underlying initial schwa /ə/ segment. This segment is only realized in the environment:]$_{\textbf{wdwd}}$[_C. Given that all words are consonant-final, the VC analysis predicts that all phrases should be consonant-final. However, as previously demonstrated, all phrases are potentially vowel-final. Under the VC analysis, phrase-final vowels are analyzed as epenthetic. The alternative analysis, under standard CV syllabification, is that all words are vowel-final, and the sequence CV$_1$]$_{\textbf{wdwd}}$[V$_2$C, is reduced to C]$_{\textbf{wdwd}}$[V$_2$C, which is the most common pattern for hiatus resolution [29].

Under the VC hypothesis, the opposition in initial phonetics between [a, i, C] is phonologically /a, i, ə/. However, in all other environments, [ə] and [i] show a complementary distribution, arguing that they are allophones of a single phoneme [30]. If a

Table of Contents for this manuscript

segmental opposition between /ə/ and /i/ is posited, then the sole environment establishing the opposition would be word-initial position, where the /ə/ is only contingently realized. Therefore, at the basic level of segmental analysis, the VC hypothesis encounters a significant problem in Kaytetye, and does not appear to offer significant advantages.

In their discussion of Arrernte, Breen and Pensalfini recognize that the VC analysis does not offer significant advantages in terms of the segmental inventory or word structure. Rather, they propose that the critical data supporting the VC analysis comes from reduplication and from Rabbit Talk, a language game. Arrernte has a number of reduplication constructions. Breen and Pensalfini analyse reduplication as a word-level phenomenon, and propose that all reduplicants have a disyllabic VC structure: VC(C)VC(C). Example (16) illustrates their analysis of Frequentative reduplication.

16) akemir-em	akemir-epir-em
get.up-PRES	get.up-FREQ-PRES
"is getting up"	"keeps getting up"

Under the VC analysis, the frequentative morpheme involves an /ep/ fixed segmentism and reduplication of the root-final syllable. Breen and Pensalfini point out that a CV syllable analysis with final vowels predicts an incorrect reduplicative form.

17)	akemire-me	*akemire-pere-me
	get.up-PRES	get.up-FREQ-PRES
	"is getting up"	"keeps getting up"

The fixed segmentism would be /pe/ and final syllable is /re/. However, an analysis parallel to Kaytetye, with CV syllables, phrasal reduplication and hiatus resolution does generate the correct reduplication construction.

18) Wd[akemire-pe]WdWd[ire-me]Wd
> Wd[akemire-p]WdWd[ire-me]Wd

Further research is required to evaluate the potential analyses of the Arrernte constructions.

5. Conclusion

We have shown that the Kaytetye partial reduplication construction is not an affixing or compounding construction. Rather it is a phrasal construction, with phonological conditioning of the reduplicant. The Kaytetye partial reduplication construction is equally problematic for theories, such as MDT, which do not posit a distinction between morphological and syntactic reduplication.

The associated motion reduplication construction, together with the Ergative allomorphy patterns provide key evidence that the shape of the minimal word in Kaytetye is VCV. The absence of word-edge consonants is typologically unusual, although within the language itself it is not unusual.

6. References

[1] Gil, D., From Repetition to Reduplication in Riau Indonesian, in Studies on Reduplication, B. Hurch, Editor. 2005.

[2] Kirchner, J.S., Minimal Reduplication. 2010, University of California, Santa Cruz: Santa Cruz.

[3] Kimper, W., Syntactic Reduplication and the Spellout of Movement Chains. 2008: University of Massachusetts, Amherst.

[4] Inkelas, S. Morphological Doubling Theory: Evidence for Morphological Doubling in Reduplication. Studies on reduplication. Berlin: Moutin de Gruyter, 2005. 65-88.

[5] Inkelas, S & C. Zoll. Reduplication: Doubling in Morphology. Cambridge University Press, 2005.

[6] Turpin, M. and A. Ross, Kaytetye to English Dictionary. 2012, Alice Springs: IAD Press.

[7] Turpin, M., A Learner's Guide to Kaytetye. 2000, Alice Springs: IAD Press.

[8] McCarthy, J. and A. Prince, Faithfulness and reduplicative identity. University of Massachusetts Occasional Papers in Linguistics, ed. L. Dickey, J. Beckman, and S. Urbanczyk. 1995.

[9] Kager, R., Optimality Theory. 1999, Cambridge: Cambridge University Press.

[10] Prince, A. and P. Smolensky, Optimality Theory: Constraint interaction in generative grammar. 2008: John Wiley & Sons.

[11] Urbanczyk, S., Reduplicative Form and the Root-Affix Asymmetry. Natural Language & Linguistic Theory, 2006. 24: p. 179-240.

[12] Keane, E., Phrasal reduplication and dual description, in Studies on Reduplication, B. Hurch, Editor. 2005. p. 239-261.

[13] Ghomeshi, J., et al., Contrastive Focus Reduplication in English. Natural Language & Linguistic Theory, 2004. 22(2): p. 307-357.

[14] Henderson, J., Topics in Eastern and Central Arrernte Grammar. 1988, University of Western Australia.

[15] McCarthy, J. and A. Prince, Prosodic Morphology: Constraint Interaction and Satisfaction. 2001.

[16] McCarthy, J. and A. Prince, Foot and Word in Prosodic Morphology: The Arabic Broken Plural. Natural Language & Linguistic Theory, 1990. 8(2): p. 209-283.

[17] Ito, J., Prosodic Minimality in Japanese. Proceedings of the Chicago Linguistics Society, 1990. 26(2): p. 213-239.

[18] Alderete, J. and K. Macmillan, Reduplication in Hawaiian: variations on a theme of minimal word. Natural Language & Linguistic Theory, 2015. 33(1): p. 1-45.

[19] Baker, B., Word Structure in Australian Languages, in The Languages and Linguistics of Australia, H. Koch and R. Nordlinger, Editors. 2014, De Gruyter Mouton.

[20] Baker, B. and M. Harvey, Word Structure in Australian Languages. Australian Journal of Linguistics, 2003. 23(1): p. 3-33.

[21] Fletcher, J. and A. Butcher, Sound Patterns of Australian Languages., in The Languages and Linguistics of Australia, H. Koch and R. Nordlinger, Editors. 2014, De Gruyter Mouton.

[22] Dixon, R.M.W., Australian Languages: their nature and development. 2002: Cambridge University Press.

[23] Turpin, M. and K. Demuth, Stress in Kaytetye, in Workshop on the Phonetic Analysis of Rhythm in Indigenous Languages. 2012: University of Auckland.

[24] San, N. and M. Turpin, Acoustic correlates of stress in Kaytetye words, in 45th Annual Conference of the Australian Linguistic Society. 2014: University of Newcastle.

[25] Koch, H., Divergent Regularity in Word-Initial Truncation in Arandic Languages, in Langauge Description, History and Development, J. Siegel, J. Lynch, and D. Eades, Editors. 2007, John Benjamins Publishing Company. p. 267-280.

[26] Koch, H., Pama-Nyungan Reflexes in the Arandic Languages, in Boundary Rider: Essays in Honour of Geoffrey O'Grady, D. Tryon and M. Walsh, Editors. 1997.

[27] Gordon, M., A perceptually-driven account of onset-sensitive stress. Natural Language & Linguistic Theory, 2005. 23(3): p. 595-653.

[28] Breen, G. and R. Pensalfini, Arrernte: A language with no syllable onsets. Linguistic inquiry, 1999. 30(1): p. 1-25.

[29] Casali, R.. Hiatus resolution. In The Blackwell companion to phonology, M.v. Oostendorp et al., Editors. 2011. Malden, Mass.: Wiley-Blackwell.

[30] San, N., Proctor, M., Turpin, M., Harvey, M., Ringbauer, K., Ross, A. & Demuth, K.. An acoustic analysis of Kaytetye vowel variability, in 46th Annual Conference of the Australian Linguistic Society (ALS), Western Sydney University. 2015.

Short vowels in L1 Aboriginal English spoken in Western Victoria

Deborah Loakes[1,2], *Janet Fletcher*[1,2], *John Hajek*[1], *Josh Clothier*[1], *Ben Volchok*[1]

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

dloakes@unimelb.edu.au

Abstract

This paper analyses the short vowel system of L1 Aboriginal English speakers from Western Victoria, as well as prelateral front vowels (in light of a merger of /el/-/æl/ in the region). The aim is to describe the short vowel system of this variety, and to explain the results of an earlier perception task carried out with the same speaker-listeners. Results show that such vowels tend to be closer (higher) and this variety of Aboriginal English has a "compressed" vowel space relative to mainstream Australian English speakers in the same area, especially with a higher /æ/ (most evident for female speakers), and a more back /ʉ:/ (restricted to male speakers). Prelateral vowels in /el/-/æl/ contexts are completely merged, showing that this phenomenon is more entrenched for Aboriginal English speakers than for the mainstream Australian English speakers in the same region.

Index Terms: Aboriginal English, vowels, prelateral merger

1. Introduction

1.1. Australian Aboriginal English

Before European colonisation in 1788, a rich variety of languages and dialects were spoken by Aboriginal people. Currently only approximately 90 remain, and only 20 are regularly spoken [1]. One of the effects of the decline in Indigenous languages is the widespread use of English by Aboriginal Australians. Latest estimates are that 83% of Aboriginal people now speak only (Australian Aboriginal) English at home [2]. Compared with mainstream Australian English, Aboriginal English is described as being characterised by differences in grammar, semantics, pragmatics, phonetics and phonology. The sound system and pronunciation of Aboriginal English is known to range from very similar to the mainstream ('light / acrolectal' Aboriginal English) to very divergent ('heavy / basilectal') with the latter having a relatively different sound system. Speakers themselves can also vary their speech depending on situation and audience [3,4].

While researchers talk about Aboriginal English as a unique variety (or group of varieties), little descriptive work has been done on the phonetic and phonological differences between the lighter acrolectal L1 Aboriginal English and mainstream Australian English. The focus of our paper is to present an initial description of the L1 Aboriginal English short vowel system from one specific region of Victoria (Warrnambool and surrounds, in the west of the state).

1.2. Aboriginal English vowels

The most comprehensive account of the phonetics/phonology of Australian Aboriginal English vowels [4,5] reports that the vowel system and vowel space can be much smaller than that found in the mainstream variety, and, more generally, that differences between Aboriginal varieties and the mainstream are less evident in more acrolectal varieties. In [5], a vowel space of female L2 Aboriginal English is presented and compared to the same vowels in mainstream Australian English. For the Aboriginal English speakers in that study, all vowels are closer (higher) than the mainstream, and the space is also more compressed in the F1 / F2 dimension, resulting in a smaller space overall.

Another study which has similar findings compares the acoustics of vowels by a group of mainstream Australian English speaking females from Katherine (NT) with English source vowels of the neighbouring Gurindji Kriol, which, with respect to grammatical features, is described as a mixed language [6]. Low vowels /æ/ and /ɐ/ are relatively close (high) compared to the local mainstream variety, and /ʉ:/ is less front. In particular, the extent of /æ/ raising is of note in Gurindji Kriol, and in that variety there is some degree of overlap between /e/-/æ/. Both [5,6] compared mainstream Australian English vowels with vowels spoken by a small number of female L2 English speakers from the Northern Territory (4 speakers in [5], 5 speakers in [6]). In [5], differences between mainstream Australian English and Aboriginal English (especially /æ/ lowering and /ʉ:/ fronting in the former) are discussed and are said to be the result of relatively recent changes that have occurred only in the mainstream variety. It appears, then, that the short vowel system, as well as the long vowel /ʉ:/, could be an important source of varietal difference.

1.3. Australian English in Western Victoria: linking production and perception

The analysis of Aboriginal English vowels in Western Victoria needs to be contextualized with respect to ongoing research on English spoken in this region. It has been previously documented that there is a vowel merger in the short vowel system where /e/ -> [æ] prelaterally [7, 8, 9]. For mainstream Australian English speakers, this merger is regionally specific to southern Victoria, occurring in Melbourne and Warrnambool (and not, for example, in Albury-Wodonga in the north). Focusing on a group from Warrnambool, we have found that the merger occurs in production and perception, and that it is "incomplete" – some speaker-listeners merge /el/-/æl/, and some keep the vowels distinct [8]. This work has also shown that production and perception are linked but not perfectly aligned, with participants who merge in production more likely to do so in perception. Additionally, there is a significant difference in how people respond to the task depending on their age, which has been analysed as a result of accent changes in the short vowel system.

This merger has been previously analysed as perceptually motivated lowering; /el/->[ɛɫ]->/æl/ [7,8]. It is phonetically interesting for many reasons, one being its relationship to the short vowel system and why it occurs in Victoria but not other regions in Australia. The decision to include Aboriginal English speakers from Warrnambool in ongoing work began in order a) to understand how other speaker groups in Victoria respond to sound changes in the mainstream accent, and b) to include Aboriginal people in the discussion of what it means to speak and perceive English in Australia. This is especially important in cases where traditional languages are no longer used, such as in Warrnambool where earlier ancestral languages have not been spoken for a number of generations, and little is known about the current speech patterns of this community.

The perception task used is a two-alternative forced-choice identification task with various 7-step vowel continua (varying in equidistant F1-F3 steps) [9]. Most important here are the control *het-hat* and *hell-Hal* conditions, which allow us to focus on how well listeners distinguish /e/-/æ/ in control and prelateral conditions. In the current study, as well as analysing the short vowel system, the production of vowels by this Aboriginal English group is investigated with a view to finding an explanation for the perceptual results reported in [9]. To date, we have seen that listeners from southern Victoria (Warrnambool, Melbourne) have trouble distinguishing /el/-/æl/ [7,8,9], while those in the northern border regions do not. The Aboriginal English listeners responded differently from mainstream Australian English listeners in the same region, with a preference for *hat* over *het* in the control condition, and a complete preference for *hell* in the merger condition [9]. While this may be in part due to lexical frequency of the items, we also need to understand the acoustics of the vowels for these speakers in production. Therefore, the aims of this paper are:

1. To describe the short vowel system of Australian Aboriginal English spoken in Western Victoria (/hVt/ vowels); and,
2. To determine whether these speakers merge /el/-/æl/ in production (/hVl/ vowels).

2. Method and Materials

2.1. Speakers and data.

Speech and perceptual data were collected by the first author in 2015 in Warrnambool and surrounding regions. 22 speakers (12 males, 10 females) with a mean age of 34 (range 19 – 65, *sd* 14.3) took part in the study. All participants in the study self-identified as Aboriginal and recognised that the study was about Aboriginal English. Speakers were sourced from two locations in Warrnambool – in the central township as well as Framlingham, which is an Aboriginal trust (once a mission) approximately 30 km from the Warrnambool city centre. Some speakers were also sourced from Heywood which is 95 km from Warrnambool city centre and approximately the same distance from Mt. Gambier. Participants included both Gunditjmara people (Warrnambool, Framlingham) and Gunditj Miring people (Heywood). Recordings were made in public spaces in the Aboriginal co-operatives in Warrnambool and Heywood, and in the health centre at Framlingham.

In all, the participants took part in a number of experimental tasks which involved questionnaires, a perception experiment (described fully in [8]) and speech recordings (wordlist and semi-spontaneous). Speech was recorded using a Zoom Handy Recorder H4n. This study focuses on the wordlist, which contained 6 repetitions of each vowel in various /hVt/ contexts (the *control* condition) and /hVl/ (the *merger* condition), and here we analyse /ɪ e æ ɐ ɔ ʊ ʉː/ for the control, and the front vowels /ɪ e æ/ prelaterally. We also include the long vowel /ʉː/ in our analysis because it is a key variable in varieties of Australian English (as discussed in 1.2). Its position in the vowel space of L1 Aboriginal English, relative to /ʊ/, is accordingly of interest. We note that the variety of Aboriginal English spoken by the participants in this study sounds (largely) impressionistically not the same as mainstream Australian English, although there is of course variation depending on the speaker.

2.2. Analysis procedure.

Data were transcribed orthographically by the fifth author. Sound and text files were automatically segmented using WebMAUS Multiple for Australian English, and a database was built for analysis in EMU/RStudio [10, 11].

In this first description of vowels by this group of speakers a static as opposed to dynamic analysis measure was used to enable comparison with earlier research on vowel spaces in Aboriginal English [5,6]. Static formant measurements at vowel targets were extracted, and the data were modelled using linear mixed effects structures in the *lme4* package in the *R* statistical environment, with separate models built for F1 and F2 per sex. For each model, we set speaker and repetition as random intercept factors. After building a maximally specified model, with PHONEME, following consonant (POSTC), AGE, and education level (EDU) as fixed factors (including all possible interactions), backward elimination of non-significant effects was performed using the R package *lmerTest*.

3. Results

3.1. Vowels in control /hVt/ context

The short vowel spaces, showing centroids derived from the mean F1/F2 measurements for the short vowels (and /ʉː/), are shown in Figures 1 (female) and 2 (male).

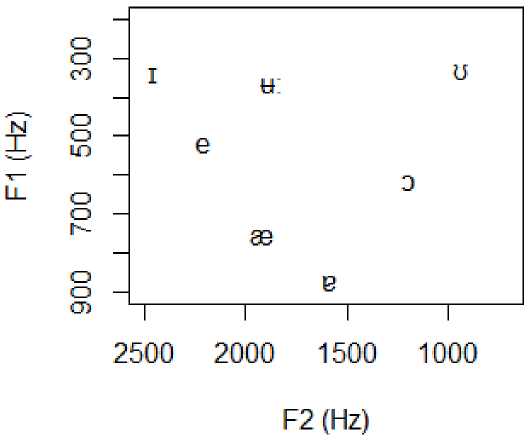


Figure 1. *Vowel targets (control, /hVt/): female speakers*

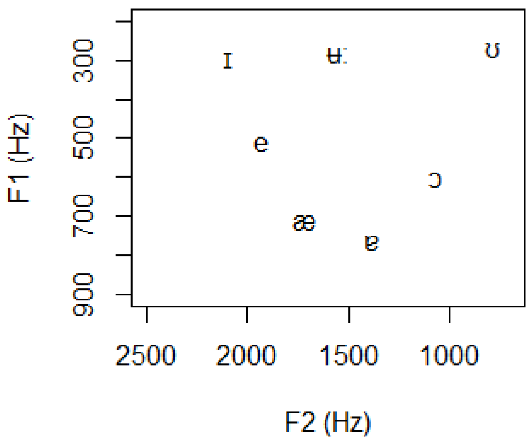


Figure 2. *Vowel targets (control, /hVt/): male speakers*

3.2. Prelateral short front /hVl/ vowels

For prelateral vowels, results are shown in Figure 3 for the female speakers, and in Figure 4 for male speakers, with ellipses for /ɪl/, /el/ and /æl/ overlaid on the centroids shown previously in Figure 1. The centroids for the prelateral words are written orthographically to better show how they correspond with the relevant control vowel.

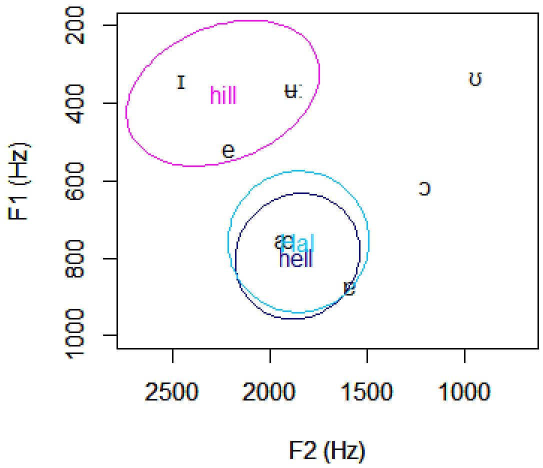


Figure 3. *Vowel targets (control) vs. front vowel targets and ellipses (prelateral): female speakers*

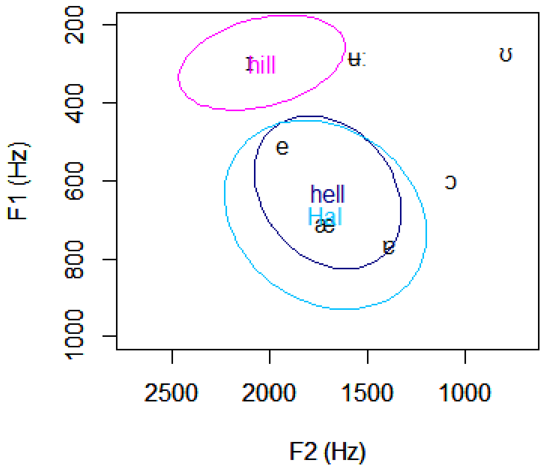


Figure 4. *Vowel targets (control) vs. front vowel targets and ellipses (prelateral): male speakers*

3.3. Data modeling.

The results of the linear mixed effects regression analyses described in § 2.2 show that the vowel phoneme, following consonant and age of the speaker were significant predictors of F1 and F2 patterns in the data. For F1 for both males and females the best fitted model had PHONEME, POSTC, and AGE as main fixed effects, with various interaction effects between: POSTC and PHONEME; POSTC and AGE; PHONEME and AGE and POSTC, PHONEME and AGE. For F2 in both sexes, the best fitting model had PHONEME, POSTC,and AGE, as main fixed effects, with interaction effects between: POSTC and PHONEME; POSTC and AGE; and POSTC, PHONEME, and AGE.

These models show an overall trend which reflect the graphical representations of the data shown in Figures 3 and 4. That is, for both males and females, there is a trend wherein vowels occurring before /l/ have higher F1 and a lower F2 than those before /t/. In other words, vowels are lowered and retracted before /l/. Confirming the visual presentation of the data (seen in 3.2), post hoc tests using a Bonferroni correction for multiple comparisons demonstrate the difference between /æ/ and /e/ prelaterally is small for males (but statistically significant, $p < .05$), and for females, prelateral /e/ has a higher F1 (is lower), than prelateral /æ/, $p < .01$.

4. Discussion and Conclusion

This paper has presented vowel spaces of a group of L1 Aboriginal English speakers in Western Victoria, with vowels in /hVt/ and /hVl/ environments. As reported for L2 Aboriginal English varieties [5,6], the L1 Aboriginal English described here also has a compressed vowel space with respect to mainstream Australian English, and most notably a higher /æ/ vowel. Compared with the most recently reported vowel spaces for mainstream Australian English [12,13], the female speakers' vowels in particular are especially compressed in the F1 dimension. For example, recent work [12,13] has shown that female speakers of mainstream Australian English tend to have a very open /æ/ vowel, with an average F1 of around 950-1100 Hz. While there is a high degree of variability for mainstream speakers, this is still a comparatively large difference when we consider the mean of around 800 Hz for Aboriginal English females (Figure 1). The Aboriginal English data also differs from recent studies with respect to the alignment of /æ-ɐ/. Because of the higher /æ/ for these

Aboriginal English speakers, the /ɐ/ vowel is at the bottom of the vowel space, rather than /æ/, which is the pattern seen for the most recent mainstream Australian English spaces [12,13]. This is in fact reminiscent of results for female speakers of the mainstream variety recorded in 1990 [14], but for both vowels those data are acoustically more open than seen for these Aboriginal English speakers by at least 100 Hz. Interestingly, F2 is less remarkable for these vowels, comparing well with [12,13].

The height of the female speakers' /e/ and /ɔ/ vowels are both acoustically similar to the Aboriginal English vowel space shown in [5], being higher than those in the mainstream variety [12,13,14]. The female Aboriginal English close vowels /ɪ ʊ ʉː/ are all (essentially) aligned in F1 (all approx 400 Hz) - which is also fairly typical for mainstream female speakers of Australian English [12]. In F2, the high vowel /ʉː/ is also practically in line with the recent data presented in [12] for Melbourne speakers. This result differs though, from Western Sydney English [13] which has a more open and more fronted realisation.

The acoustic space in Figure 3 (males) cannot be compared to other data for Aboriginal English which have focused on female speakers only, but it can still be compared with mainstream Australian English data for males [12,13,14] and considered with respect to the female speakers. In this case, the vowel space is only slightly compressed compared to mainstream Australian English. In the F1 dimension, /æ/ produced by the Aboriginal English speakers is only slightly higher than seen in other recent work [12,13] and is very close to the data presented in [14] (recorded in the 1990s). Like for the female Aboriginal English speakers, the F2 values for /æ-ɐ/ are within the range for the mainstream accent [12,13] and are reminiscent of F1 values reported in [14]. Additionally, each of /ɪ e ɔ ʊ/ fall within the range shown in [12,13] for the mainstream accent, in both F1 and F2; /ʉː/ however, is more back than any of the reports on mainstream Australian English [12,13,14]. This backed /ʉː/ mirrors what other researchers comparing Aboriginal English with mainstream Australian English have found for female speakers [5,6], but is different from what we observed for the female speakers in the same region whose /ʉː/ patterns with results observed for mainstream speakers from Melbourne ([12] as discussed above).

For the prelateral context (Figure 3, 4), it is clear that /e/ is most affected by coarticulation. This is also evident when looking at the centroids; while /ɪ/ and /æ/ are essentially the same in /hVt/ and /hVl/ environments, there is a complete merger of F1 and F2 for *hell-Hal*, so that the centroids (and ellipses in the case of the female speakers) are entirely overlapping. While there are trends in the data as far as this overlap of *hell* and *hat* is concerned, it is (as discussed in 3.3) only significant for males in F1, and females in F2.

Comparing these new production results with what we know about perception from the same participants [9], the reasons for particular responses are more clear. In particular, Aboriginal English listeners chose /æ/ quite "early" in the *het-hat* continuum (at Step 2 of 7), which is not surprising given we now know they also have an especially high /e/ and /æ/ in their own production (esp. females). This is similar to some findings for (primarily older) listeners of mainstream Australian English from the same region [8] who also tend to choose /æ/ relatively early in the continuum. We know now, too, that the Aboriginal English speakers overlap (merge) prelateral /el/ with /æl/ in production. This is different from mainstream Australian English listeners, who have a wider variety of production behaviour. Some mainstream speakers merge and some do not, and there is therefore greater variability in ellipses, and also in patterns of responses in perception [8,9]. In listening, the Aboriginal English participants always preferred *hell* across the continuum (there was no crossover from *Hal->hell* at all for these listeners). Given what we know about production from the current study, we might infer that prelateral /el/-/æl/ stimuli would be completely ambiguous in perception for these listeners and, along with *hell* being lexically more frequent than the competing *Hal*, it is understandable why /el/ was preferred across the board.

This study has gone some of the way towards understanding how L1 Aboriginal English speakers produce short vowels (plus /ʉː/), and why they respond in unique ways in perception. Future work will focus on vowel trajectories, in /hVl/ environments, to better understand coarticulation in this context. We are also interested in sociophonetic patterning of vowels for this group, especially in light of the differing behaviour for males and females where /ʉː/ and /æ/ are concerned.

5. References

[1] McConvell, P., & N. Thieberger (2001). State of Indigenous languages in Australia – 2001. Canberra: Department of the Environment and Heritage.

[2] Australian Bureau of Statistics (2006). Population distribution, Aboriginal and Torres Strait Islander Australians. (cat. no. 4713.0). Accessed online January, 2015. Available at: http://www.abs.gov.au/ausstats/abs@.nsf/mf/4705.0.

[3] Eades, D. (2000). Aboriginal English (Pen Note 93). Newtown, NSW: Primary English Teaching Association.

[4] Butcher, A. (2008). Linguistic aspects of Australian Aboriginal English. Clinical Linguistics and Phonetics, 22(8): 625-642.

[5] Butcher, A. and V. Anderson (2008) The vowels of Australian Aboriginal English. In Interspeech 2008, available http://www2.hawaii.edu/~vanderso/Butcher-Anderson.pdf

[6] Jones, C. F. Meakins & H. Buchan. 2011. Comparing vowels in Gurindji Kriol and Katherine English. AJL. 31, (3). 305-326.

[7] Loakes, D., J. Hajek. J. Clothier, J. Fletcher. 2014. Identifying /el/-/æl/: a comparison between two regional Australian towns. Proceedings of the 15th SST, Canterbury: ASSTA, pp.41-44

[8] D. Loakes, J. Clothier, J. Hajek, J. Fletcher. 2014. An investigation of the /el/-/æl/ merger in Australian English" AJL. 34 (4), 436-452.

[9] Loakes, D., J. Fletcher, J. Hajek & J. Clothier. 2016. What reaction times reveal about listener groups: L1 Aboriginal English and Standard Australian English responses to a prelateral merger-in-progress. LabPhon16, Ithaca, NY.

[10] Kisler, T., Schiel, F. & Sloetjes, H., 2012. Signal processing via web services: WebMAUS. Hamburg, Germany, 30-34.

[11] R Development Core Team. 2008. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.Rproject.org.

[12] Cox, F., S. Palethorpe, K. Miles and B. Davies (2014) "Is there evidence for region specific vowel variation in /hVd/ word list data from AusTalk?" paper presented at the ALS Annual Conference. Newcastle University, Newcastle, Dec 10.

[13] J. Elvin, D. Williams and P. Escudero. 2016. Dynamic acoustic properties of monophthongs and diphthongs in Western Sydney Australian English. JASA, 140 (1). 576-581.

[14] Cox, F., 2006. The Acoustic Characteristics of /hVd/ Vowels in the Speech of Some Australian Teenagers. AJL, (26), 2. 147-179.

Tailoring phonetic learning to the needs of individuals on the basis of language aptitude

Mark Antoniou and Melissa Blair

The MARCS Institute for Brain, Behaviour and Development, Western Sydney University
m.antoniou@westernsydney.edu.au, 17647700@student.westernsydney.edu.au

Abstract

Many learners have difficulty discerning non-native speech sounds. The present study sought to tailor training to the needs of learners to improve learning. Participants learned five artificial languages, containing word pairs that differed on a single, critical phonetic feature. Previous research established that language aptitude predicted learning for these artificial languages. Therefore, we divided subjects into high and low aptitude groups and modified training in several ways. High aptitude subjects benefitted from numerous training methods, whereas low aptitude subjects only benefitted from increased exposure. These results are useful for identifying which training types benefit high and low aptitude learners.

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

1. Introduction

Many learners commonly experience difficulty when learning the sounds of a non-native language. It is well documented that non-native talkers speak with a foreign accent. Studies on cross-language speech perception have demonstrated that listeners also hear with an 'accent' [1], meaning that non-native 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 perception may lead to subsequent difficulties in language acquisition, such as vocabulary, comprehension, and literacy. Thus, by improving phonetic learning, it may be possible to ultimately improve language learning more generally. Although it is widely accepted that learners differ from one another, the majority of language teaching materials have assumed that all individuals learn in the same way. This consequently results in a mismatch in variation between learners in classrooms, and the homogeneity implied by the one-size-fits-all approach of most language course books. The aim of the present study was to examine the interaction between individual differences and training paradigm design.

For language learning to be successful, it is necessary to strike a balance between a complex combination of internal and external factors [4]. Internal factors refer to characteristics of the individual learner, such as cognitive abilities and motivation. External factors, on the other hand, refer to characteristics of the learning situation, including the language to be learned, and the characteristics of the teaching method, such as whether corrective feedback is provided. Recent work shows that the interaction between subject-internal and -external factors determines how successful learning will be. For example, talker variability (i.e., exposing learners to the speech of numerous talkers) is thought to result in more robust learning outcomes, better generalisation to novel stimuli, and greater long-term retention of phonetic information [5]. However, not all learners benefit from talker variability. Perrachione et al. [6] asked American-English native speakers 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. The implication of these findings is that it is crucial to take both individual differences and training methods (internal and external factors) into consideration when prescribing speech training.

Although foreign language learning can be challenging, research has identified several factors that are considered to be advantageous for non-native learners. First, it has been proposed that closely related languages may be easier to learn than others. This may be due to knowledge transfer from a learner's native language [7], or cross-language phonetic similarity [3]. Second, it has been suggested that certain language features might be easier to acquire for all learners regardless of past language experience [8]. Third, individual differences in cognitive abilities, such as working memory [9] and language aptitude [10] may contribute to language learning, and explain the variability that is typically observed between learners. The central idea in foreign language aptitude research is that the underlying reason why learners vary considerably in foreign language learning is because some possess a specific talent for learning languages. The concept of language aptitude is controversial, and raises questions concerning whether it is an innate ability or responsive to training, and whether it predicts language learning success and thus might be used to tailor training [11].

The present study is the latest in a series examining the interaction between subject-internal and -external factors in the learning of non-native phonetic contrasts.

Antoniou et al. [12] compared English monolinguals to Mandarin-English and Korean-English bilinguals in their ability to learn novel words from three artificial languages containing phonetic contrasts similar to those that occur in English, Mandarin, and Korean. Results indicated that all groups learned the Mandarin-like language most successfully, followed by the English-like, whereas the Korean-like language was the most difficult to learn. For the difficult Korean-like language, Korean-English bilinguals showed an advantage, suggesting that similarity to the native language may aid the learning of 'difficult' non-native phonetic contrasts.

Antoniou and Low [13] extended this work by presenting five artificial languages to English monolinguals and also measured the learners' pre-training cognitive profiles (language aptitude and working memory). Results revealed that, like in Antoniou et al. [12], the Mandarin-like language was the easiest to learn, followed by the English-like, whilst the Korean-like, Vietnamese-like, and Arabic-like languages were all learned equally poorly. In terms of individual differences, language aptitude reliably predicted successful learning in four of the five languages, and may therefore be a useful variable for separating good from poor learners when prescribing training.

The aim of the present study was to tailor training proactively to maximise learning outcomes based on the needs of individual learners. Based on the findings of [13], subjects were separated into high and low aptitude groups by administering a standardised assessment of language aptitude prior to the commencement of training. Training was modified in three ways (one for each of the three difficult to learn languages from [13]). The first training modification doubled the number of exposure trials to provide learners with more opportunity to learn the critical phonetic contrast without increasing other demands. Second, minimal pair words were chunked, that is, they were presented in couplets with the aim of highlighting the critical phonetic contrast that differentiates the words. Third, during the exposure phase, subjects were quizzed and corrective feedback was provided. These different training types were expected to benefit the high and low aptitude groups in different ways. Specifically, we hypothesized that:

1. High aptitude learners will outperform low aptitude learners.

2. High aptitude learners will benefit from multiple training types, and will show the largest benefit from corrective feedback.

3. Low aptitude learners may need more exposure to learn nonnative phonetic contrasts.

4. Low aptitude learners may be overwhelmed by feedback, especially for hard-to-learn languages.

2. Method

2.1. Participants

Twenty-eight Australian English native speakers (*M* age = 22.1 years; *SD* = 5.3) took part in the study. All were undergraduate Psychology students at Western Sydney University. None reported any history of neurological deficits. All passed an air conduction audiogram at 25 dB HL at 500, 1000, 2000, and 4000 Hz.

2.2. Language aptitude assessment

Prior to training, subjects' language aptitude was assessed using the LLAMA standardised tests [14]. Specifically, LLAMA subtest B was used to measure vocabulary learning ability. In the LLAMA B test, 20 pictures are presented on-screen, and subjects are given two minutes to click on each picture and learn to associate the visually presented novel word that corresponds with each picture. An audible *ding* is presented as feedback for a correct response, and a *beep* is presented for incorrect responses. LLAMA B scores range from 0-100 in increments of 5, with higher scores indicating greater language aptitude. Subjects were divided into two groups based on their LLAMA B scores, creating high (>50) and low aptitude (≤50) groups (see Table 1).

Table 1. *High and low aptitude group sample sizes, mean ages, and LLAMA B scores.*

Group	*n*	Age	LLAMA B (*SD*)
High aptitude	14	23.7	68.9 (9.0)
Low aptitude	14	20.5	35.0 (13.4)

2.3. Stimulus materials

Each language was comprised of four pairs of consonant-vowel monosyllables, and ended with vowels /e/, /i/, /o/, and /u/. The word pairs differed on a single, critical phonetic feature that was non-native, but similar to that of a phonetic feature which occurs in a natural language (see Table 2). The English-like artificial language contained a bilabial fricative voicing contrast /ɸ/-/β/, phonetically similar to that of the English labiodental fricative voicing contrast /f/-/v/, but differing in place of articulation (bilabial as opposed to labiodental). The English-like stimuli were produced by a phonetically trained male English speaker. The Mandarin-like language contained a dental-retroflex stop contrast /t/-/ʈ/, phonetically similar to the dental-retroflex sibilant contrast found in Mandarin /s/-/ʂ/, but realised with a different manner. As the retroflex versus non-retroflex stop contrast is native to Gujarati, the stimuli were produced by a female Gujarati native speaker. The Korean-like artificial language contained a voiceless fricative lenition contrast /θ/-/θˈ/, phonetically similar to the Korean voiceless stop lenition contrast /t/-/t*/, but differing in manner. The Korean-like stimuli were produced by a phonetically trained female native speaker of Korean. The Arabic-like language differentiated words using a voiceless velar-uvular ejective contrast /k'/-/q'/, similar to the native Arabic voiceless velar-uvular plosive contrast /k/-/q/, but produced with a different airstream mechanism. The Arabic-like stimuli were produced by a male native speaker of Quechua, a language containing velar and uvular ejectives. The Vietnamese-like language differentiated words using velar voiced plosive and implosive stops /g/-/ɠ/, similar to the bilabial voiced plosive and implosive stop contrast that occurs in Vietnamese /b/-/ɓ/, but with a different place of articulation. The Vietnamese-like stimuli were produced by a female native speaker of Sindhi, a language that contrasts velar voiced plosive and implosive stops.

The stimulus recordings were conducted inside a sound attenuated booth using a Shure SM58 cardioid microphone attached to a boom stand. The duration of the vowels were normalised to 350 ms for the Korean-like, Vietnamese-like, Arabic-like and Mandarin-like languages, and to 300 ms for the English-like language. The same vowel was spliced into both constants within a minimal pair to ensure that subjects were distinguishing the words based solely on the critical consonant distinction.

2.4. Procedure

The artificial language learning experiment was presented using Sennheiser HD 280 Pro headphones connected to a HP Pro Book 650 laptop running E-Prime software. Stimulus output level was calibrated to 72 dB SPL. The presentation order of the five artificial languages was counter-balanced across subjects.

A passive exposure training paradigm was adapted from past research [12], [13]. All training methods began with an exposure phase, which paired a picture and sound onscreen.

Table 2. *Natural and artificial language contrasts and their frequency of occurrence in the world's languages.*

Artificial language	Natural language contrast	Occurrence in world's languages (%)	Artificial language contrast
English-like	/f/-/v/ Labiodental **fricative voicing**	50.8	/ɸ/-/β/ Bilabial **fricative voicing**
Mandarin-like	/s/-/ʂ/ **Dental-retroflex** sibilants	20.2	/t/-/ʈ/ **Dental-retroflex** stops
Arabic-like	/k/ - /q/ **Voiceless velar-uvular** plosives	20.0	/k'/-/q'/ **Voiceless velar-uvular** ejectives
Vietnamese-like	/b/-/ɓ/ Bilabial **voiced plosive-voiced implosive stops**	10.0	/g/-/ɠ/ Velar **voiced plosive-voiced implosive stops**
Korean-like	/t/-/t*/ Voiceless alveolar stop **lenition**	1.8	/θ/-/θ'/ Voiceless dental fricative **lenition**

The exposure phase consisted of eight words (four pairs) and 12 repetitions, giving 96 trials in total. The number of exposure trials were doubled to 192 in the double exposure training type (used for the Arabic-like language). Words were presented in random order at a rate of every 3.5 seconds with no requirement for response to proceed to the next word. In the chunking training type (used for the Vietnamese-like and Mandarin-like languages), minimal pair words were presented in couplets (e.g., /gu/ followed by /ɠu/). In the feedback training type (used for the Korean-like and English-like languages), following exposure to each couplet, subjects were quizzed by having both pictures presented onscreen and then selecting the correct picture following playback of each of the two words, and corrective feedback was provided.

The exposure phase was immediately followed by the test phase, which was identical for all three training types. In the test phase, subjects were presented with eight repetitions of each word, giving 64 test trials in total per language. Each word was auditorily presented and all eight corresponding pictures were displayed on the screen. Participants were required to match the heard word to the correct picture by pressing the corresponding number (1-8) on a keyboard. The test was self-paced and no corrective feedback was provided.

3. Results

Word identification scores for each of the five artificial languages are depicted in Figure 1. Initial inspection of Figure 1 suggests that both groups found the Mandarin-like language easiest to learn. The high aptitude group appears to have a general learning advantage over the low aptitude group, with the exception of the Vietnamese-like language, which both groups learned poorly.

To determine how language aptitude relates to training paradigm design, we conducted a 2 (Group) × (5 Languages) factorial ANOVA. A main effect of group, $F(1, 26) = 16.4$, $p < .001$, $\eta_p^2 = .386$, revealed that the high aptitude group exhibited an overall learning advantage over the low aptitude group. Further, there was a main effect of language, $F(4, 104) = 39.5$, $p < .001$, $\eta_p^2 = .603$, as well as a Group × Language interaction, $F(4, 104) = 3.1$, $p = .02$, $\eta_p^2 = .106$. To analyse the interaction, a series of posthoc *t*-tests were conducted. These posthoc analyses confirmed that the high aptitude group outperformed the low aptitude group in learning of the Mandarin-like language, $t(26) = 2.82$, $p = .009$, the Arabic-like language, $t(26) = 3.75$, $p = .001$, and the English-like language, $t(26) = 3.07$, $p = .005$. The high and low aptitude groups did not differ in their learning of the Vietnamese-like, $t(26) = 0.53$, $p = .604$, or Korean-like languages, $t(26) = 2.36$, $p = .026$ (note that Bonferroni correction requires an adjusted alpha level of .05 / 5 = .01).

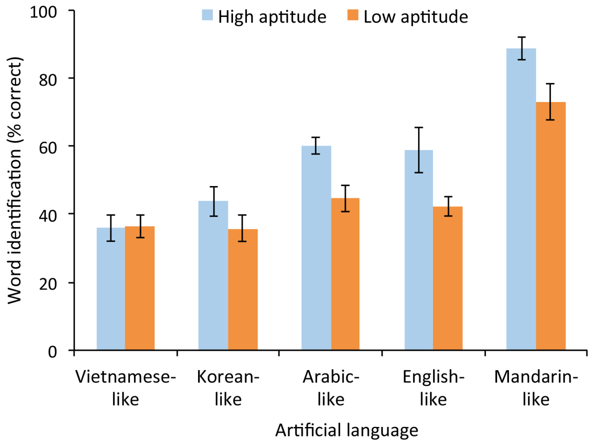


Figure 1. *High and low aptitude groups' mean word identification scores for each artificial language. Error bars depict standard error of the mean.*

By comparing the present results to those of Antoniou and Low, we may evaluate the effectiveness of the three training types relative to the baseline passive exposure condition used in their study (i.e., 96 exposure trials, no chunking, no feedback). Recall that in that study, the Vietnamese-like, Korean-like, and Arabic-like languages were learned equally poorly (~40%). Also bear in mind that Antoniou and Low did not separate good from poor learners. Our double exposure training improved learning of the Arabic-like language for both the high and low aptitude groups (grand average = 52.4%). The low aptitude group benefitted from chunking for the easy-to-learn Mandarin-like language. However, low aptitude learners did not benefit from chunking or feedback for the other languages. The high aptitude group benefitted from all three training types (double exposure for the Arabic-like language, chunking for the Mandarin-like, and feedback for the Korean-like).

4. Discussion

The present study investigated the interaction between language aptitude and training paradigm design in phonetic learning. As predicted, high aptitude learners showed a general learning advantage over those with low aptitude. High aptitude learners also benefitted from a variety of training types. In contrast, low aptitude learners benefitted from doubled exposure for the difficult-to-learn Arabic-like language. They did not benefit from either chunking or feedback for the difficult-to-learn languages, and learned the Korean-like and Vietnamese-like languages equally poorly.

These findings are consistent with a cognitive resource limitation view of speech perception and phonetic learning. Such an account is offered by the active control model [15], according to which closed-loop processing routines called active control structures monitor incoming speech stimuli in a context sensitive way. Crucially, this active control draws from a finite pool of cognitive resources and thus when the task facing the learner becomes more difficult, more cognitive resources are required, and perceptual performance will decline when these resources are depleted. Support for the active control model comes from studies investigating speech perception under different levels of cognitive load [16]. If speech processing depends on active control structures which in turn depend on the availability of cognitive resources then performance costs should be exacerbated under cognitively demanding conditions. Therefore, it stands to reason that low-aptitude learners might be overwhelmed by a phonetic learning task in which they are faced with the challenge of learning a difficult non-native phonetic contrast. This resource limitation hinders their ability to attend to the relevant information in the contrasts being learned.

High aptitude learners benefitted from several training types. First, they benefitted from chunking of minimal pair stimuli when learning the Mandarin-like language. Second, they benefitted from corrective feedback for the English-like language. Third, they benefitted from increased passive exposure to the Arabic-like language. These results suggest that for all of the above languages, high aptitude learners possess sufficient cognitive resources in reserve to take advantage of the additional information presented in each training paradigm.

In contrast, low aptitude learners do not have sufficient cognitive resources in reserve and thus are unable to attend to the additional information presented during training (e.g., chunking or feedback). They did, however, benefit from increased exposure to the Arabic-like language, probably because this training method does not increase processing demands.

The findings are consistent with [12], [13]. As in those studies, English listeners found the Mandarin-like language easiest to learn, whereas the Vietnamese-like and Korean-like languages were the most difficult to learn. However, our training modifications resulted in improved learning outcomes relative to these previous studies. This improvement was most pronounced for the Arabic-like language. The results demonstrate that cognitive abilities interact with language learning outcomes. Specifically, they support the assertion that language aptitude is a useful variable for tailoring training [11]. Thus, we strongly advocate training individuals differently when it comes to phonetic learning.

5. Conclusions

We divided learners into high and low language aptitude groups and exposed them to several different training methods. High aptitude subjects benefitted from numerous training methods, whereas low aptitude subjects only benefitted from increased exposure. In sum, the present findings demonstrate that when it comes to language learning, one size does not fit all. Our results have implications for speech training paradigms.

6. References

[1] M. Antoniou, M. D. Tyler, and C. T. Best, "Two ways to listen: Do L2-dominant bilinguals perceive stop voicing according to language mode?," *J Phon*, vol. 40, no. 4, pp. 582–594, Jul. 2012.

[2] M. Antoniou, C. T. Best, and M. D. Tyler, "Focusing the lens of language experience: Perception of Ma'di stops by Greek and English bilinguals and monolinguals," *J Acoust Soc Am*, vol. 133, no. 4, pp. 2397–2411, Apr. 2013.

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

[4] M. Antoniou, M. Ettlinger, and P. C. M. Wong, "Complexity, training paradigm design, and the contribution of memory subsystems to grammar learning," *PLoS ONE*, vol. 11, no. 7, p. e0158812, Jul. 2016.

[5] J. S. Logan, S. E. Lively, and D. B. Pisoni, "Training Japanese listeners to identify English /r/ and /l/: A first report," *J Acoust Soc Am*, vol. 89, no. 2, pp. 874–886, Feb. 1991.

[6] T. K. Perrachione, J. Lee, L. Y. Y. Ha, and P. C. M. Wong, "Learning a novel phonological contrast depends on interactions between individual differences and training paradigm design," *J Acoust Soc Am*, vol. 130, no. 1, pp. 461–472, 2011.

[7] R. C. Major, "Transfer in second language phonology: A review," in *Phonology and second language acquisition*, vol. 36, J. G. Hansen Edwards and M. L. Zampini, Eds. Amsterdam: John Benjamins, 2008, pp. 63–94.

[8] F. R. Eckman, "Typological markedness and second language phonology," in *Phonology and second language acquisition*, vol. 36, J. G. Hansen Edwards and M. L. Zampini, Eds. Amsterdam: John Benjamins, 2008, pp. 95–115.

[9] A. Baddeley, "Working memory and language: An overview," *J Comm Dis*, vol. 36, no. 3, pp. 189–208, May 2003.

[10] J. B. Carroll, "Foreign language proficiency levels attained by language majors near graduation from college," *Foreign Lang Ann*, vol. 1, no. 2, pp. 131–151, Dec. 1967.

[11] Z. Dörnyei and P. Skehan, "Individual differences in second language learning," in *The handbook of second language acquisition*, C. J. Doughty and M. H. Long, Eds. 2003, pp. 589–630.

[12] M. Antoniou, E. Liang, M. Ettlinger, and P. C. M. Wong, "The bilingual advantage in phonetic learning," *Biling Lang Cogn*, vol. 18, no. 4, pp. 683–695, Oct. 2015.

[13] M. Antoniou and T. Low, "Individual differences in language learning," presented at the CoEDL Fest, Western Sydney University, 2016.

[14] P. Meara, "Llama, Language Aptitude Tests: The manual," *Swansea: Lognostics*, 2005.

[15] J. S. Magnuson and H. C. Nusbaum, "Acoustic differences, listener expectations, and the perceptual accommodation of talker variability," *J Exp Psychol Hum Percept Perform*, vol. 33, no. 2, pp. 391–409, 2007.

[16] M. Antoniou and P. C. M. Wong, "Poor phonetic perceivers are affected by cognitive load when resolving talker variability," *J Acoust Soc Am*, vol. 138, no. 2, pp. 571–574, Aug. 2015.

L2 phonological category formation and discrimination in learners varying in L2 experience

Mona M. Faris[1], *Catherine T. Best*[2], *& Michael D. Tyler*[3]

[1]The MARCS Institute, Western Sydney University, Australia
[2]School of Humanities and Communication Arts, Western Sydney University, Australia
[3]School of Social Sciences and Psychology, Western Sydney University, Australia
m.faris@westernsydney.edu.au, c.best@westernsydney.edu.au, m.tyler@westernsydney.edu.au

Abstract

Non-native phones that are perceived as speech-like, but do not closely resemble any single first-language (L1) category, are assimilated as uncategorised. The Perceptual Assimilation Model for Second-Language (L2) Speech Learning [1] predicts that such phones are likely to be acquired as L2 categories, which should result in improvements in discrimination. This study investigated the acquisition of uncategorised L2 phones and discrimination performance in L1 Egyptian-Arabic learners varying in L2 Australian-English experience. While no firm conclusions can yet be drawn about L2 category formation, age of acquisition and L2 vocabulary size predicted discrimination accuracy, but this was dependent upon assimilation type.

Index Terms: vowel perception, L2 immersion, individual variability

1. Introduction

Research on L2 speech perception has demonstrated that, unlike young children, adult learners experience difficulty in discriminating and acquiring certain L2 phones [2]. It is well known, for example, that the discrimination of L2 phones varies as a function of both L1 attunement and the specific contrasts tested. For instance, German and Spanish learners of English differ on their discrimination of the English vowels /ɛ/ and /æ/ [3]. The Spanish listeners were able to successfully discriminate between these two English vowels, which they perceived as two contrasting L1 vowel categories (Spanish /e/ and /a/, respectively). The German listeners, on the other hand, discriminated this English contrast poorly, which they may have perceived as instances of a single L1 vowel category (German /ɛ/ or /ɛː/). Similarly, Japanese learners experience varying degrees of perceptual difficulty in discriminating between certain Australian-English vowel contrasts [4]. Good discrimination was observed for the L2 contrast /iː/-/ɪ/, which they perceived as two separate L1 vowel categories, namely, /iː/ and /i/, respectively. The L2 contrast /iː/-/ɪə/ was discriminated poorly, however, since they perceived both phones within the contrast as the single L1 phoneme /iː/.

Several models on L2 speech perception have been developed to account for the variations in discrimination performance on L2 contrasts by L2 learners. One of the most prominent models of L2 speech acquisition is the Perceptual Assimilation Model for L2 Speech Learning [PAM-L2; 1]. PAM-L2 predicts the likelihood of establishing new L2 categories, which would result in changes in discrimination performance over L2 learning time. According to PAM-L2, L1 attunement shapes L2 speech acquisition in ways that imply a shared L1-L2 phonological system. PAM-L2 makes various predictions about L2 speech learning based on the way in which L2 phones are initially assimilated to the L1 phonological system. An L2 phone that is perceived as somewhat similar to an existing L1 phoneme will be assimilated as categorised and may vary in its goodness-of-fit to the native ideal. If, however, it is perceived as speech-like, but does not closely resemble any of the L1 categories, then it will be assimilated within the L1 phonological space as uncategorised. However, phones that are perceived as non-speech will not be assimilated within the phonological space. When considered as pairs of phones, various predictions about L2 learning are possible. Consider assimilation types where one or both phones are uncategorised, namely, Uncategorised-Categorised and Uncategorised-Uncategorised assimilations. According to PAM-L2, a new L2 category is likely to be formed for phones assimilated as uncategorised. This is predicted to result in improvements in discrimination over L2 learning time. It is the discrimination and acquisition of those contrasts that are of interest in the present study.

New L2 category formation is influenced not only by the different ways in which pairs of L2 phones assimilate to the L1 phonological system, but there are also a number of factors that have been shown to affect L2 speech acquisition. High discrimination accuracy is associated with a longer length of residence [LOR; 5], a younger age of acquisition [AOA; 6] and age of immersion in the L2 speaking environment [AOI; 7], use of the L2 is relatively more than the L1 [proportion L2 use; 2], a higher L2 vocabulary size [L2 VS; 4], and a longer period of L2 training prior to L2 immersion [EFL; 8]. Given the variability among L2 learners, the effect of these factors on L2 perception will be considered in the current study. While the PAM-L2 predictions are based on beginner learners who are immersed in the L2 speaking environment, its principles are also applicable to other learning situations.

This study is part of a larger project that aims to test the PAM-L2 predictions of new L2 phonological category acquisition and to track the changes in discrimination performance over L2 learning time in learners varying in L2 experience for contrasts assimilated as Uncategorised-Categorised and Uncategorised-Uncategorised. Here we present data from the first testing session of the longitudinal study to investigate how variations in L2 experience influence L2 category acquisition and discrimination, and to test PAM contrast assimilation predictions. The PAM-L2 predictions were examined in Egyptian-Arabic (EA) learners of Australian-English (AusE) who had been exposed to the L2 prior to immersion. Based on the perceptual assimilation results in [9], two Uncategorised-Categorised (/ʊ/-/ɔ/, /ɪ/-/e/) and seven Uncategorised-Uncategorised (/ɪə/-/iː/, /æɪ/-/ɑe/, /əʉ/-/ʉː/, /oː/-/ʉː/, /oː/-/əʉ/, /æ/-/ɐ/, /eː/-/ɜː/) AusE vowel contrasts were selected for the current study.

2. Method

2.1. Participants

Thirty-eight native adult EA speakers participated in the study (20 females, M_{age} = 41 years, age range: 17 – 73 years). They were recruited from the Greater Western Sydney community and through snowball sampling. Participants varied to some extent on each of the six factors (see Table 1). They indicated that they were native-born speakers of EA, with no hearing or language impairments, and normal or corrected-to-normal vision. None of the participants had had any extended stay in an English-speaking country prior to immersion in Australia. English instruction in Egypt was typically from non-native speakers of British or American English. They received monetary reimbursement for their participation.

Table 1. *Characteristics (means and ranges) of the learners on each of the six factors.*

Variable	Mean	Range	
		Min.	Max.
Length of residence	1.41 years	0.03 years	6.34 years
Age of acquisition	13 y/o	2 y/o	52 y/o
Age of immersion	40 y/o	16 y/o	71 y/o
Proportion of L2 use	0.37	0.06	0.61
L2 vocabulary size	9200 words	4600 words	14200 words
English as a foreign language	10.03 years	0 years	23 years

2.2. Stimulus and Apparatus

The stimuli were the same as those used in [9]. Briefly, the auditory stimuli were produced in a sound-attenuated booth at the Western Sydney University by two female speakers of AusE (34 and 44 years old) recruited from the Greater Western Sydney region. All AusE monophthongs /ɐ, e, ɪ, ɔ, ʊ, æ, ɐː, eː, iː, oː, ʉː, ɜː/, diphthongs /ɑe, æɔ, æɪ, əʉ, ɪə, oɪ/, and /ə/ [10] were produced in /ˈhVbə/ nonsense words. Selected tokens were those produced with a falling intonation and spoken with a consistent speaking rate across talkers. The tokens containing the vowels /æ, ɐ, ɪ, e, ʊ, ɔ, eː, ɜː, iː, oː, ʉː, æɪ, ɑe, ɪə, əʉ/ were selected for the current study.

The stimuli were recorded at a 44.1 kHz sampling rate using a Shure SM10A headset microphone connected to an Edirol UA-25EX external USB sound card. The recordings were high-pass filtered at 70 Hz to attenuate low-frequency noise and to correct for the DC component. Tokens were ramped such that the onset and offset of each token had a 10 ms fade-in and 20 ms fade-out, respectively. Four tokens per vowel category were selected from both speakers resulting in a total of 120 tokens (15 vowels x 4 repetitions x 2 speakers). Any audible clicks detected in the tokens were excised.

L2 vocabulary size was assessed using a bilingual version of the Nation and Beglar L2-English Vocabulary Size Test [11]. It is an assessment of decontextualised knowledge of written receptive vocabulary presented in a multiple-choice format. As the English version of the test requires grammatical knowledge and fair reading abilities, a bilingual version of the test was developed by the first author. The bilingual version of the test required participants to select one out of four translated definitions that best match the test word or phrase. They were given one of two equivalent versions of the test, each containing 100 multiple-choice questions. The readability of the test and the accuracy of the translations were checked by native EA speakers prior to the administration of the test.

2.3. Procedure

Participants first completed an AXB categorial discrimination task for each of the nine AusE vowel contrasts. Participants indicated whether the vowel in the middle token (X) belonged to the same phonemic category as the vowel in either the first (A) or last (B) token. To encourage phonological perception, all three tokens per trial were physically different, with tokens A and B produced by a different speaker than token X. The interstimulus interval was 1 s. Participants were asked to attend to the first vowel in the nonsense syllable and to select one of two keys on a computer keyboard. If a response was not collected within 2 s, the trial was repeated a random number of trials later. No feedback was provided. Participants were first familiarised with the procedure on three practice trials with feedback, and the tokens were produced by a different female AusE speaker than those from the experimental trials. For each AXB task, there were 64 trials, which were randomised for each participant. All four trial types (i.e., AAB, ABB, BAA, BBA) were presented an equal number of times per contrast. As there were three tokens per speaker, using a Latin-square design, each token was presented an equal number of times in each position (i.e., A, X, B). The order of presentation of each AXB contrast was pseudorandomised.

Participants then completed an L1 perceptual assimilation task with goodness-of-fit ratings. On a given trial, they were presented with a /ˈhVbə/ nonsense syllable over headphones and were instructed to attend to the target vowel. A grid was then presented containing all L1 core phonemic (/a, i, u, aː, iː, uː, eː, oː, aw, aj/) and allophonic ([æ, æː, ɑ, ɑː, ɛː, ɛ̆̽ː, e, o, ɪ, ɪː, ʊ, ʊː, ə]) vowel categories, and /ʔ/ presented in Arabic CVC or CV keywords, with the vowels highlighted in red. Using a computer mouse, participants selected an L1 keyword containing the vowel closest to the auditorily presented AusE vowel. After the token was presented again, they rated its goodness-of-fit to their chosen EA vowel using a scale from 1 (strange) to 7 (perfect). No feedback was provided. A keyword selection and rating response were required to be made within 6 s and 3.5 s, respectively, otherwise the entire trial was reinserted into the random sequence. There were 120 trials (15 vowels x 2 speakers x 4 repetitions), the intertrial interval was 500 ms, and the presentation order of the trials was randomised for each participant.

In addition to the L1 perceptual assimilation task, an L2 perceptual assimilation task was administered in order to allow for inferences to be made about new L2 phonological category formation. The procedure was similar to that of the L1 task except that participants categorised the L2 vowels to L2 AusE vowel category labels. All 18 AusE vowels were presented in CVC or CV English keywords, with the vowels highlighted in red. The order of presentation of the two perceptual assimilation tasks was counterbalanced. Stimulus presentation and response collection for the AXB task and both perceptual assimilation tasks were controlled using PsyScope X B57 on a MacBook laptop, Sennheiser HD 650 headphones, and an Edirol UA-25EX external USB sound card.

Participants were given the vocabulary size test, and a language background information questionnaire in order to collect information on the participants' AOA, AOI, LOR, EFL, and proportion of L2 usage.

3. Results

3.1.1. *New L2 phonological category formation*

Inferences about new L2 phonological category formation were made based on the perceptual assimilation patterns from both L1 and L2 perceptual assimilation tasks. For a given L2 phone, if it was uncategorised in the L1 but categorised in the L2, then this was taken as indirect evidence that a new L2 phonological category had been formed. Phones were deemed categorised if an L1 category label was consistently selected more than 50% of the time, otherwise it was deemed uncategorised. No systematic differences were found in whether an AusE vowel was categorised to an L1 core phonemic versus allophonic category, so the allophonic vowel categories were collapsed into the appropriate main phonemic categories [see 9]. The only two AusE vowels that were categorised to an L1 category were /æ/ and /ɐ/, which were categorised to the EA /aː/ 53% and 54% of the time, respectively, while none of the AusE vowels were categorised to an L2 vowel category label.

The individual perceptual assimilation patterns revealed a high degree of variability in terms of whether a given L2 phone was categorised or uncategorised, both in the L1 and L2 tasks. Given the high degree of interindividual variability, for each individual participant instances were identified where an L2 AusE vowel was uncategorised in the L1, but categorised in the L2. A binomial logistic regression was conducted with each of the six variables (i.e., AOA, AOI, LOR, L2 VS, proportion of L2 usage, and EFL) to determine whether any of those factors are related to the likelihood of forming a new L2 phonological category. An L2 phone was uncategorised in the L1 but categorised in the L2 in 21% of instances. The logistic regression was not statistically significant, $\chi^2(6) = 3.339$, $p >$ .05, suggesting that none of the six factors reliably predicted the likelihood of new L2 phonological category acquisition for this first testing session of the longitudinal project.

3.1.2. *Assessing PAM's predictions of discrimination*

PAM assimilation types were determined in the same way as in [12]. When the L2 phones were considered as contrasts, there was a high degree of interindividual variability in the PAM assimilation patterns. For example, while the contrast /oː/-/əʉ/ was assimilated as Uncategorised-Uncategorised in the L1 perceptual assimilation task at the group level, individual participants assimilated it either as Uncategorised-Uncategorised or Uncategorised-Categorised. Given this high degree of variability, to analyse the discrimination results, we adopted the same approach as in [12]. Specifically, each individual's assimilation type for each of the nine contrasts was determined. The mean discrimination accuracy scores were then grouped according to individual assimilation type rather than on vowel contrast. For example, the discrimination accuracy scores for Uncategorised-Uncategorised assimilations were grouped together, regardless of the contrast in which they occurred. Individual assimilation patterns were determined per individual for both the L1 and L2 tasks.

Individual assimilation types were compared across the L1 and L2. There were cases where a contrast was assimilated as a Single-Category in the L1, but as a Two-Category contrast in the L2, suggesting that the participant was able to discern phonological differences between the pair of contrasting L2 phones, *and* that they had learned the new L2 contrast. Therefore, it may be more meaningful to consider both L1 and L2 perceptual assimilation patterns than either one alone. Taking into account both L1 and L2 assimilation types, we created a composite L1-L2 assimilation type by selecting the L1 or L2 assimilation type that was predicted to result in the more accurate discrimination across the two. For instance, for an individual participant, if a given contrast was assimilated as Uncategorised-Uncategorised in the L1, but as a Two-Category in the L2, then the L2 perceptual assimilation type was selected. Similarly, if a contrast was Uncategorised-Categorised in the L1, but Single-Category in the L2, then the L1 perceptual assimilation type was selected. Eight percent of cases were Single-Category assimilations, another 8% were Two-Category, 40% were Uncategorised-Categorised, and 42% were Uncategorised-Uncategorised. Since only 1% of cases were of Category-Goodness assimilations (comprised of two data points), they were excluded from further analyses.

A one-way between-subjects analysis of variance was conducted to determine if the discrimination accuracy scores vary as a function of assimilation type. There was a significant difference in the discrimination accuracy scores among the assimilation types, $F(3, 336) = 5.446$, $p = .001$. A Bonferroni post-hoc comparison revealed Two-Category assimilations were discriminated more accurately than Single-Category, $M_{\text{diff}} = 15.92\%$, $p = .001$, $SE = 4.27\%$. Uncategorised-Categorised assimilations were discriminated more accurately than Single-Category assimilations, $M_{\text{diff}} = 10.68\%$, $p = .009$, $SE = 3.34\%$. The results are displayed in Figure 1.

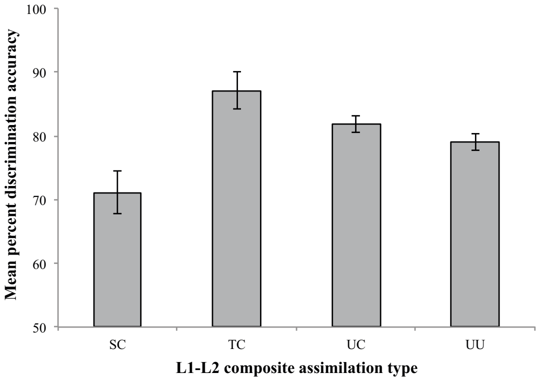


Figure 1: *Mean percent discrimination accuracy for the composite L1-L2 assimilations. Error bars represent standard error of the mean*

3.1.3. *Effect of the six factors on discrimination accuracy*

The relationship between the six predictors on discrimination accuracy was assessed using the composite L1-L2 assimilation types. Bivariate Pearson correlations were conducted between the mean percent discrimination accuracy score for a given assimilation type with each of the six factors. The correlations are presented in Table 2, with significant bivariate Pearson correlation coefficients presented in bold, which ranged from .202 to -.490. A younger age of acquisition was associated with more accurate discrimination for all assimilation types, except for Single-Category. Uncategorised-Categorised and Uncategorised-Uncategorised contrast assimilations each yielded a larger number of significant correlations than Single-Category and Two-Category assimilations combined.

To determine whether any of the factors predicted discrimination performance, a separate standard multiple regression was conducted for each composite L1-L2 assimilation type. A high L2 vocabulary size predicted poorer discrimination accuracy for Single-Category assimilations (i.e., a negative correlation), $F(6, 21) = 4.114$, $p = .007$, and accounted for approximately 47% of the variance ($R^2 = .540$, adjusted $R^2 = .409$). A younger age of acquisition was a significant predictor of better discrimination accuracy for Uncategorised-Categorised assimilations, $F(6, 131) = 10.90$, $p < .001$, and accounted for approximately 32% of the variance ($R^2 = .333$, adjusted $R^2 = .303$). Similarly, a younger age of acquisition significantly predicted higher discrimination accuracy scores for Uncategorised-Uncategorised assimilations, $F(6, 138) = 2.560$, $p = .022$, and accounted for approximately 8% of the variance ($R^2 = .100$, adjusted $R^2 = .061$). None of the factors significantly predicted discrimination accuracy for Two-Category assimilations.

Table 2. *Bivariate Pearson correlations between the mean percent discrimination accuracy scores for each composite L1-L2 assimilation type with each of the six factors.*

Composite L1-L2	Mean discrimination accuracy	LOR	AOI	AOA	EFL	Prop. L2 use	L2 VS
SC	71	-.122	**-.407***	.001	.221	-.077	**-.476***
TC	87	-.296	-.142	**-.490****	.164	.352	.231
UC	82	**-.376****	-.146	**-.480****	**.445****	**.202***	**.364****
UU	79	**-.205***	-.123	**-.266****	**.248****	.036	.116

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

4. Discussion

This study aimed to examine new L2 category formation and discrimination accuracy in learners varying in L2 experience. Vowel perception was shown to be highly variable [e.g., 12]. Despite this variability, PAM's predictions of discrimination were upheld such that both Two-Category and Uncategorised-Categorised assimilations were discriminated more accurately than Single-Category assimilations.

By accounting for variability within individual participants who differed on factors related to L2 experience, we have shown that, to some extent, differences in L2 discrimination accuracy may be explained by such factors. A high L2 vocabulary size predicted poor discrimination accuracy. According to PAM-L2, a steadily expanding L2 vocabulary size is beneficial for L2 learners as it forces them to attend to important phonetic details in the L2 that are not employed in the L1, and in turn, help learners distinguish between minimally contrasting L2 words. But, a rapidly expanding L2 vocabulary may be detrimental for L2 learners as it may cause them to fossilise, or settle on a suboptimal common L1-L2 phonological category, thus curtailing further L2 development. At this initial stage of testing, the learners' L2 vocabulary size was high, averaging 9200 words. The current study did not assess the rate of L2 vocabulary acquisition. However, the acquisition of English vocabulary and grammar are normally the key focuses of L2 acquisition in schools and universities in Cairo. It may be tentatively inferred that the L2 vocabulary was acquired rapidly prior to L2 immersion. As this study forms part of a larger longitudinal study, there will be an opportunity to track how changes in L2 vocabulary size affect discrimination accuracy over L2 learning time. Vocabulary size for an average native English speaker is roughly 20,000 words [13], so there remains room for vocabulary expansion.

While discrimination accuracy was affected to some extent by some of the factors, none of the factors reliably predicted discrimination accuracy for Two-Category assimilations. It is unsurprising given that it is L1 attunement that helps the listener distinguish between phones assimilated as Two-Category, which is consistent with PAM's framework.

Individual differences may also play a role in new L2 phonological category formation. The results revealed that none of the factors significantly predicted the likelihood of new L2 phonological category formation. However, as only 21% of cases were of an L2 phone that was uncategorised in the L1 but categorised in the L2, there may not be sufficient statistical power to detect those influences. Consequently, no firm conclusions may be made at this stage of the longitudinal study. The effect of the six factors on category formation will be examined longitudinally as a function of changes in L2 immersion experience. This should in turn be reflected in changes in discrimination performance over L2 learning time.

The next stage of this project will be to examine the developmental changes over a 12-month period of L2 immersion by tracking changes in discrimination performance as a function of perceptual assimilation, and how discrimination performance is affected by the six factors.

5. References

[1] C.T. Best and M.D. Tyler, Nonnative and second-language speech perception: Commonalities and complementarities, in Second language speech learning: The role of language experience in speech perception and production, M.J. Munro and O.-S. Bohn, Eds. 2007, John Benjamins: Amsterdam. p. 13-34.

[2] J.E. Flege and I.R.A. MacKay, "Perceiving vowels in a second language," *Stud. Second Lang. Acquis.*, vol. 26, pp. 1-34, 2004**.**

[3] J.E. Flege, O.S. Bohn and S. Jang, "Effects of experience on non-native speakers' production and perception of English vowels," *J. Phonetics*, vol. 25, no. 4, pp. 437-470, 1997.

[4] R. Bundgaard-Nielsen, C.T. Best and M.D. Tyler, "Vocabulary size is associated with second-language vowel perception performance in adult learners," *Stud. Second Lang. Acquis.*, vol. 33, pp. 433-461, 2011.

[5] J.E. Flege and S. Liu, "The effect of experience on adults' acquisition of a second language," *Stud. Second Lang. Acquis.*, vol. 23, no. 4, pp. 527-552, 2001.

[6] G. Jia, W. Strange, Y. Wu, J. Collado and Q. Guan, "Perception and production of English vowels by Mandarin speakers: Age-related differences vary with amount of L2 exposure," *J. Acoust. Soc. Am.*, vol. 119, no. 2, pp. 1118-1130, 2006.

[7] W. Baker, P. Trofimovich, M. Mack and J. E. Flege, "The effect of perceived phonetic similarity on non-native sound learning by children and adults," in BUCLD, Boston, MA, vol. 26, no. 1, pp. 36-47, 2002.

[8] J. Cebrian, "Experience and the use of non-native duration in L2 vowel categorization," *J. Phonetics*, vol. 34, no. 3, pp. 372-387, 2006.

[9] M.M. Faris, C.T. Best and M.D. Tyler, "An examination of the different ways that non-native phones may be perceptually assimilated as uncategorized," *JASA*, vol. 139, no. 1, pp. EL1-EL5, 2016.

[10] F. Cox and S. Palethorpe, "Australian English," *J. Int. Phonetic Assoc.*, vol. 37, pp. 341-350, 2007.

[11] P. Nation and D. A. Beglar, "A vocabulary size test," *Lang. Teacher*, vol. 31, pp. 9-13, 2007.

[12] M.D. Tyler, C.T. Best, A. Faber and A.G. Levitt, "Perceptual assimilation and discrimination of non-native vowel contrasts," *Phonetica*, vol. 71, pp. 4-21, 2014.

[13] P. Nation and R. Waring, Vocabulary size, text coverage and word lists, in Vocabulary: Description, Acquisition and Pedagogy, N. Schmitt and M. McCarthy, Eds. 1997, Cambridge University Press: Cambridge, pp. 6-19.

The Effects of Order and Intensity of Training on the Perception and Production of English /e/-/æ/ by Cantonese ESL learners

Janice Wing Sze Wong

Hong Kong Baptist University
janicewong@hkbu.edu.hk

Abstract

This study investigated the effects of the order of provision of perception and production training and the intensity of the two training types on the perception and production of English /e/-/æ/ by Cantonese ESL learners. Thirty-five subjects were assigned into four groups with different training order and intensity levels. Results showed that all groups improved their perception and production of /e/-/æ/, but the groups which received production training before perceptual training performed better than those who were trained perceptually first. Training intensity, i.e. how many days the training sessions were spread over, however, did not affect the degree of improvement.

Index Terms: High Variability Phonetic Training, explicit articulatory training, training order, training intensity

1. Introduction

Research studies conducted in the last few decades have shown that difficulties in learning L2 contrasts can be ameliorated by laboratory training such as perceptual-only training (e.g. [1]-[2]), production-only training (e.g. [3],[4]), or a combination of both (e.g. [5],[6]). These studies have shown inconsistent results although most of them reported some improvement of learning in either/both modalities. The present study aimed to evaluate the effectiveness of both perception and production training on the perception and production of the English /e/-/æ/ contrast among Hong Kong Cantonese ESL learners. This vowel pair is commonly-confused by this group of learners [7]. Also, it would be useful in practice to examine the effects of training order, i.e. training in one modality prior to the other and the reverse, as previous research (e.g. [5],[6]) had not regarded this as a variable.

Meanwhile, training intensity had also been assumed as a constant in training studies and the consensus was that intensive training could contribute to the learning of these difficult contrasts. Thus, how intensive a treatment should be so as to achieve optimal results, or whether different intensity levels could generate diverse training effects, have been overlooked. No training studies on L2 speech perception and production have systematically investigated the effect of training intensity on subjects' performance; only in recent decades had training intensity begun to receive some attention in research on young children with language delays and disabilities and their results were not consistent (e.g. [8],[9]).

Thus, besides examining the effects of training in both modalities and the order of provision of training on the perception and production of /e/-/æ/ contrast, the effects of training intensity were also addressed in this study in order to garner some insights in ways to optimize training effects and provide pedagogical insights to teachers and learners.

2. Methodology

2.1. Design

All the participants took part in these three phases:

PHASE 1. Pretest Phase: including one production pretest and one perception pretest;

PHASE 2. Treatment Phase: depending on the training intensity (standard, one session per day for 20 days vs. intensive, five sessions per day for four days) and order (entire HVPT paradigm completed first followed by production training, vs. the reverse)

a. *High Variability Phonetic Training*: 10 training sessions were offered, each lasted for 10 minutes.

b. *Explicit Articulation Training*: 10 training sessions were offered, each lasted for 10 minutes.

PHASE 3. Posttest Phase: including one production posttest, one perception post-test, and three perception Tests of Generalization (TG1, TG2 and TG3).

2.2. Participants

A total of 35 secondary school students were recruited to the present experiment. They had Hong Kong Cantonese as their L1 and English as the L2 and were aged around 16 to 17. They all started learning English as an L2 at the age of 3.32 (SD = .53) for an average of 13.6 years (SD = .82). No one had resided in any English-speaking countries. They all reported that they had no hearing or speaking impairment.

All the subjects received the same number of training sessions and finished training in one modality first before another. They were hence divided into four groups depending on the training order and intensity:

- Group 1 – HPS: 9 received 10 **H**VPT perception training sessions, followed by 10 explicit articulatory **p**roduction training sessions and they completed one session per day for 20 days (**s**tandard training);
- Group 2 – HPI: 9 received also **H**VPT first before the **p**roduction training, but they received five training sessions per day for four days (**i**ntensive training);
- Group 3 – PHS: 9 subjects received the **p**roduction training first before **H**VPT and they received one session per day for 20 days (**s**tandard training);
- Group 4 – PHI: 8 received also the **p**roduction training first before **H**VPT but with five training sessions per day for four days (**i**ntensive training).

A total of eight native General American (GA) English speakers (6 female and 2 male) were also invited to produce stimuli for the tests and training. Their ages ranged from 35 to

40.

2.3. Setting and Apparatus

All the subjects completed the tests and training sessions in a language laboratory. They completed all perceptual training and test sessions by using a computer program designed by the researcher. They listened to the audio tokens presented in the program and completed an identification task. All the data were saved into a Microsoft Access database for analysis.

In the explicit production training sessions, videos in which a native GA speaker displayed and taught the articulation of vowels were given to the subjects to watch before practicing the target vowels with the researcher. The subjects recorded the production test tokens using Adobe Audition 1.5 through a Shure SM58 microphone.

2.4. HVPT Stimuli

Six of the eight native GA English speakers produced stimuli used in both the perceptual pre/posttest and the HVPT training. All of them produced 20 minimal word pairs, i.e. a total of 40 tokens. All words were CVC monosyllabic words with different onsets and codas. One of the speakers, i.e. a familiar speaker to the subjects, recorded also a new word list with 20 /e/-/æ/ minimal word pairs for the use in TG2 (new words by a familiar speaker). Another speaker whose voice had not appeared in the training or tests recorded the same set of tokens for the use in TG3 (familiar words by a new speaker). The last speaker who had not recorded any tokens for the training stimuli or the tests, i.e. a new speaker, recorded another new list with 20 /e/-/æ/ minimal word pairs for the use in TG1 (new words by a new speaker). All the minimal pairs in TGs were with various CVC contexts and syllable structures (mono-, di- and poly-syllabic) with a view to testing the transfer of learning under various conditions.

2.5. Production Training Materials

A female native GA speaker whose voice had not appeared in any perceptual stimuli recorded the training items. The video recordings were made in a soundproof room and the face of the speaker was put against a blue background. A Canon EOS 600D digital camera with video recording function was used. Full HD 1080p video recording at 25 frames were made. Audio was recorded using Adobe Audition 1.5 through reading into a Shure SM58 microphone with the X2u XLR-to-USB adapter for digital audio recording and was synced to the video afterwards.

2.6. Procedure

2.6.1. Pretest Phase

All groups participated in both the production and perception tests in the first phase. The production pretest was administered first to avoid subjects' cueing or being exposed to the items which would appear later in the perception pretest.

- *Production Pretest:* The subjects were given a word list of 20 words (with 10 /e/ and 10 /æ/) and had to record all the words which would appear either in the perception pretest or the training. To ensure authentic performance and that the subjects could produce also other segments apart from the vowel, before the pretest, the subjects could hear the pronunciations of the words produced by a native speaker who had not been involved in the study. The instructions for this production pretest were offered to the subjects in the form of five practice trials and they had to produce them with natural loudness and speaking rate. They were not provided with any audio prompts or instructions during the recording. They could also pause and resume during the recording based on their own pace. The test took less than five minutes to complete.
- *Perception Pretest:* The subjects could get access to the computer program to complete the identification test which included 50 questions (40 words with either /e/ or /æ/ with 10 distractors). They did 10 practice trials before the test, which were not analyzed. Each stimulus could be played by the subjects as many times as they needed before they chose the answer from three choices with conventional English orthography, or a blank for a free answer, in which they could type their own word. The frequency of occurrence of the correct answer that appeared in the four serial positions, i.e., word 1, word 2, word 3, free answer, were equal; thus the chance level was 25%. This design was an attempt to avoid using simply two choices with 50% of chance level. The program was also designed to limit the subjects from not answering one question before moving onto another. The whole test could be completed within 10 minutes.

2.6.2. Treatment Phase

Both HPS and HPI groups participated in HVPT first before the production training whereas PHS and PHI groups received the production training before HVPT. Details about the two types of training in this phase were as follows:

- *High Variability Phonetic Training:* A total of 40 stimuli (20 /e/ and 20 /æ/) produced by six different native English speakers, all randomized in terms of speakers and word order in each session, were presented to the subjects. The subjects were trained on a two-alternative forced choice identification task which directed their attention to identifying the target word and raised the training effect, unlike the four choices used in the pretest. During training, immediate feedback was given; at the end of each session, their total scores were also shown.
- *Explicit Production Training:* The subjects were provided with videos in which a native speaker demonstrated the articulation of the vowels. The subjects had to watch the video first and read after the video host. A word list with 20 different words containing one of two target vowels was given to the subjects and they pronounced after the researcher at least three times and immediate corrective feedback was given. Articulation information of the vowel pairs, i.e. the tongue position, vowel openness, as well as the length of the vowels was also emphasized in the each session explicitly with pictures as illustrations and mirrors to help them better produce the target sounds.

The "I" groups and "S" groups also differed in terms of the number of training sessions received per day. The "I" groups received five training sessions per day and in between each session, a short break with refreshments was given; the "S" groups received one training session per day. Thus, the "I" groups would complete all the training within 4 days whereas the "S" groups spent 20 days to finish the training.

2.6.3. Posttest Phase

The Posttest Phase involved one production posttest (same as the Production Pretest) which was completed before the four perception posttests (posttest, TG1, TG2 and TG3), which were all done on the same day.

- *Production Posttest:* same as the Production Pretest
- *Perception Posttest:* same as the Perception Pretest
- *Test of Generalization 1:* The subjects heard 40 tokens (with 20 /e/ and 20 /æ/) spoken by a new speaker whose voice was not heard in any of the training stimuli or the tests. The procedures were similar to those administered in the Perception Pretest, and subjects were also given four choices to choose from.
- *Test of Generalization 2:* The subjects had to listen to 40 new words (with 20 /e/ and 20 /æ/) spoken by a familiar speaker, who had been one of the speakers in the training stimuli. Procedures were the same as those in TG1.
- *Test of Generalization 3:* The subjects revisited 40 familiar words which they had come across in the perception training sessions, but the words were produced by a new speaker. Again, the procedures were the same as those in TG1 and TG2.

2.7. Evaluation of Production Data

The production scores were evaluated by directly counting the number of accurate productions. The productions of the subjects were transcribed twice by a phonetically-trained researcher for whom Cantonese was the L1 and English the L2. The intra-rater reliability obtained was 92.30% (α = .833). Another researcher who had English as L1 also transcribed the data phonetically. During transcription, they transcribed phonetically the word they heard, which was not limited to only the target vowels. The reliability check was done without referring to any completed transcriptions. The inter-rater reliability was 91.36% (α = .815). A follow-up acoustic analysis on half of the productions, by checking the F1 and F2, F3 values and the vowel durations, was conducted by a third phonetically-trained researcher to confirm that the transcriptions aligned with the acoustic measures and were reliable. The acoustic analysis results were consistent with the transcription.

3. Results

3.1. Perceptual Performance

3.1.1. *Effects of training: Pretest vs. Posttest*

The following boxplot shows the results of the four groups in the pretest and posttest:

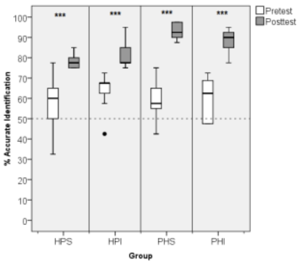


Figure 1: *Mean percentages of correct identification of the four groups in pre (white) & posttest (dark) [*** = p < .001].*

A four-way repeated measures ANOVA was computed using Test (pretest, posttest), Training Order (HP vs PH), Training Intensity (Standard vs. Intensive) and Vowel (/e/, /æ/) as factors. It showed significant main effects of Test [$F(1,31) = 182.94$, $p < .001$], Training Order [$F(1,31) = 5.30$, $p = .028$] and Vowel [$F(1,31) = 5.95$, $p = .021$]; yet, Training Intensity was not a significant factor ($p = .618$). The interactions Test × Order [$F(1,31) = 9.10$, $p = .005$], Test × Training Intensity [$F(1,31) = 7.44$, $p = .008$], Vowel × Training Intensity [$F(1,31) = 9.95$, $p = .004$] and Test × Vowel × Training Intensity [$F(1,31) = 4.93$, $p = .034$] were all significant. Planned comparisons with Bonferroni correction on Test × Training Order interaction showed that all groups improved significantly from pretest to posttest (both at $p < 0.001$). A significant difference between Training Order in the posttest, but not in the pretest ($p = 1.00$), was also found. In the posttest, PH groups outperformed HP groups by 10.61% ($p < .001$).

3.1.2. *Generalizability of training*

The following three boxplots show the results in TG1, TG2 and TG 3 (from left to right):

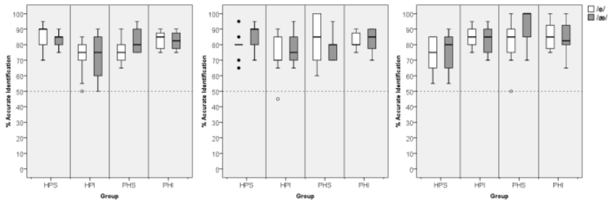


Figure 2: *Mean percentages of correct identification of vowels in all TG1, TG2 and TG3 (from left to right) across groups.*

For TG1 (new words produced by a new speaker), a three-way ANOVA with Training Intensity, Training Order and Vowel was computed. The result showed no significant main effect, but only one interaction, Training Intensity × Training Order, was robust [$F(1,31) = 10.80$, $p = .003$]. This was due to the fact that subjects trained under the PH order performed significantly better than the HP order in both intensity levels. A simple test of effects showed that only under intensive training condition did those subjects who were trained under PH perform better than HP for 11.15% ($p = .003$).

While for TG2 (new words produced by a familiar speaker), another three-way ANOVA showed that no main effect and interaction were significant. Yet, the figures still showed that the subjects could identify 79.43% and 81% accurately for the vowels /e/ and /æ/ respectively, meaning that the groups performed similarly.

With regard to TG3 (familiar words produced by a new speaker), the same ANOVA showed only a significant main effect of Training Order [$F(1,31) = 5.22$, $p = .029$] since the PH groups outperformed HP groups by 6.53%

Perceptual learning was shown to be able to generalize to new speakers and new tokens, and providing production training first before HVPT appeared to be more useful.

3.2. Production Performance: Effects of training (Pretest vs. Posttest)

A four-way repeated measures ANOVA was computed using Test (pretest, posttest), Training Order (HP vs PH), Training Intensity (Standard vs. Intensive) and Vowel (/e/, /æ/) as factors. Only the main effects of Test [$F(1,31) = 275.45$, $p < .001$] and Training Order [$F(1,31) = 4.80$, $p = .036$] were robust, indicating that all groups showed improvements in production accuracy after training and the order of training played a role in the learning. The interaction Test × Training Order [$F(1,31) = 11.58$, $p = .002$] was also significant.

Planned comparisons with Bonferroni correction showed that those subjects who were trained under the production training before HVPT outperformed the other two groups by 11.49% in terms of production accuracy ($p < .001$) while in the pretest their performance were similar ($p = .511$). However, neither the main effects of Vowel ($p = .531$) and Training Intensity ($p = .265$) nor the other interaction effects, were significant. This boxplot displays the results of production pretest versus posttest across groups:

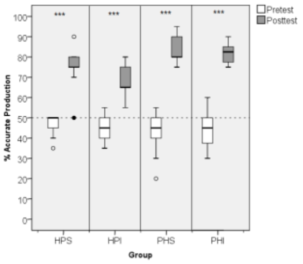


Figure 3: *Percentage of target production of the four groups between the pretest (white) & posttest (dark) [*** = p < .001].*

4. Discussion

The present results indicated that providing both high-variability phonetic training and explicit articulatory training could benefit L2 leaners in the perception and production of non-native vowel contrasts /e/-/æ/. The success could be attributed to the use of high variability stimuli in HVPT and the corrective feedback given in the production training. It is not clear, however, which training or exactly what elements in the training benefited the participants more, although this is not the goal of the present research. Still, it is found that the order of training plays a rather important role in determining the degree of success.

Training in production training before perception training helped the participants improve more in both the perception and production of the target vowel pair than the reverse order of training. This finding is intriguing as previous studies utilizing training in both modalities have not considered this as a factor. Modifying the articulatory patterns in production training first before being exposed to a wide variety of acoustic cues from the perceptual stimuli was more helpful in both perception and production. This appears to suggest that there exist some links between perception and production and perceptual representations may be heavily articulatorily-based [10]. Changes in the basic perceptual unit would lay an even more solid foundation for the tuning of perceptual representations as well as phonological-motor mapping when the learners were to receive further perceptual training later, than those who were trained in perception before production. Although the present study did not aim to test any theories, the findings suggested a crucial role played by articulatory gestures in the perception and production of non-native speech sounds. Yet, the present study has not traced the performance progress during training to gauge the amount of benefit brought by each of the two types of training or how training in one modality first can benefit the other. Further research that directly compares the effectiveness of perception-only and production-only on the two modalities would also be useful.

Another finding in this study was that training intensity did not show any significant effects in the learning in the two modalities. One plausible explanation was that the "active ingredients" of the training may matter more [11]. Intensity alone is insufficient to determine the training outcome; rather, it is the active ingredients (e.g. the type of training utilized, the training components used, how the training was delivered, etc.) that contribute to the learning. Following this line of reasoning, it is probably the adoption of high-variability perceptual stimuli, identification tasks, the use of corrective feedback in articulatory training, etc. that already become the contributing parameters leading to successful learning, lowering the possible effect that might be brought by training intensity. The optimal number of training that can improve learners' performance also merits more investigation.

5. Conclusions

The present study showed that training order had an effect on the perception and production of the English vowel pair /e/-/æ/ while training intensity did not. It suggested that when some training was provided, learning would occur and would not be affected by how the training sessions were spread over a period of time. It was rather the training order that influenced the training outcome. This finding may benefit second language learners who have difficulties in these non-native contrasts to train themselves at their own pace. Future research can investigate the effects of a wider variety of intensity levels and its interaction with other vowel training programs.

6. References

[1] Bradlow, A., Pisoni, D., Akahane-Yamada, R., and Tohkura, Y., "Training Japanese listeners to identify English /r/ and /l/: IV. Some effects of perceptual learning on speech production", *J. Acoust. Soc. Am.* 101:2299-2310, 1997.

[2] Iverson, P., and Evans, B.G., "Learning English vowels with different first language vowel systems II: Auditory training for native Spanish and German speakers", *J. Acoust. Soc. Am.*, 126:866-877, 2009.

[3] Kartushina, N., Hervais-Adelman, A., Frauenfelder, U. H., and Golestani, N., "The effect of phonetic production training with visual feedback on the perception and production of foreign speech sounds," *J. Acoust. Soc. Am.*, 138(2):817-832, 2015.

[4] Saloranta, A., Tamminen, H., Alku, P., and Peltola, M. S., "Learning of a non-native vowel through instructed production training", *Proc. ICPhS2015*, Paper 0235, 2015.

[5] Wong, J. W. S., "The Effects of Training Diversity in Training the Perception and Production of English Vowels /ɪ/ and /iː/ by Cantonese ESL learners", *Proc. Interspeech2013*, 2113-2117, 2013.

[6] Aliaga-Garcıa, C., and Mora, J. C., "Assessing the effects of phonetic training on L2 sound perception and production," in M. A. Watkins, A. S. Rauber, and B. O. Baptista [Eds], *Recent Res. Sec. Lang. Phon./Phono.: Percep. Produ.*, 2-31, Cambridge Scholars Publishing 2009.

[7] Chan, A. Y. W., and Li, D. C. S., "English and Cantonese Phonology in Contrast: Explaining Cantonese ESL Learners' English Pronunciation Problems", *Lang. Cul. Curri.*, 13:67-85, 2000.

[8] Fey, M.E., Warren, S.F., Brady, N., Finestack, L.H., Bredin-Oja, S. L., Fairchild, M., Sokol, S., and Yoder, P. J., "Early effects of responsivity education/prelinguistic milieu teaching for children with developmental delays and their parents", *J. Speech, Lang., Hear. Res.*, 49:526-547, 2006.

[9] Gray, S., "Word-learning by preschoolers with specific language impairment: Effect of phonological or semantic cues," *J. Speech, Lang., Hear. Res.*, 48:1452-1467, 2005.

[10] Best, C. T., "A direct realist view of cross-language speech perception: New Directions in Research and Theory", in Winifred Strange [Ed], *Speech perception and linguistic experience: Theoretical and methodological issues*, 171-204, York Press, 1999.

[11] Baker, E., "Optimal intervention intensity", in *Inter. J. Speech-Lang. Path.*, 14:401-409, 2012.

The Australian SpIN™ speech in noise test

Peter J. Blamey[1,2], *Maryam Zargarbashi*[1,2], *Jeremy K. Blamey*[1,2], *and Elaine Saunders*[1,2]

[1] Blamey and Saunders Hearing Pty Ltd, Australia
[2] Department of Audiology and Speech Pathology, The University of Melbourne, Australia
[3] Faculty of Science Engineering and Technology, Swinburne University of Technology

peter.blamey@blameysaunders.com.au, mzargarbashi@student.unimelb.edu.au,
jeremy.blamey@blameysaunders.com.au, elaine.saunders@blameysaunders.com.au

Abstract

An Australian recording of the QuickSIN sentences was tested with 12 normally hearing adults and 12 adults with impaired hearing to determine the signal-to-noise ratio at which 50% of words were recognised (SNR50). Speech was presented from the front and four-talker babble was presented from behind the listener. Eight of the 12 lists gave equivalent SNR50 values (–3.8 to –4.8 dB) for listeners with normal hearing. SNR50 for listeners with impaired hearing averaged 10.3 dB unaided and 2.7 dB in the aided condition, indicating an average 7.6 dB benefit from the use of hearing aids with adaptive directional microphones.

Index Terms: speech, noise, signal-to-noise-ratio, hearing loss, hearing aid, directional microphone

1. Introduction

A problem commonly reported by hearing aid (HA) users is difficulty understanding speech in noise [1]. In 1970, Carhart and Tillman [2] emphasized the importance of measuring speech recognition in the presence of background noise, suggesting speech-in-noise tests be included in the standard audiometric test battery. Several tests have been developed for this purpose, including the HINT (Hearing In Noise Test) [3] and the QuickSIN (Quick Speech In Noise) Test [4]. In 1978, Plomp [5] proposed a model of hearing impairment comprised of two independent components: an attenuation component and a distortion component which was equivalent to a signal-to-noise-ratio (SNR) component. In keeping with Plomp's model, the QuickSIN estimates an SNR50 value which is the SNR at which a listener recognises 50% of words in sentences correctly and compares it with the SNR50 for normally hearing people to estimate the SNRloss [4]. If Plomp's model is correct then the SNRloss and the audiogram together will provide a better characterization of an individual's hearing loss than either measure on its own.

The QuickSIN can also be used to estimate the SNR50 with HAs fitted to find the HA benefit as the difference between the aided and unaided SNR50 values. It is common these days for digital HAs to provide substantial benefits in background noise as a consequence of the use of directional microphones. Depending on the directions of the speech and noise, the benefits can be as great as 8 deciBels (dB) increase in the effective SNR [6]. Being able to experience this effect and demonstrate the size of the benefit in background noise can be an important factor in convincing people of the benefit of wearing HAs.

The original QuickSIN test materials [7][8] were recorded with an American speaker. Twelve equivalent QuickSIN sentence lists were developed. Each QuickSIN list consisted of six Institute of Electrical and Electronics Engineers (IEEE) phonetically balanced sentences, to provide subtle semantic cues with strong syntax cues [9]. Each sentence contained five key words presented at a fixed level in multi-talker babble. The multi-talker babble level was increased by 5 dB for successive sentences to produce SNR values from 25 dB to 0 dB to accommodate the performance of normal to severely hearing-impaired individuals [7][8]. Killion et al. [4] found normal hearing participants' SNR50 was +2 dB SNR and used this value to calculate SNRloss. The American QuickSIN has been widely used, but is not ideal for Australian listeners because of the American accent of the speaker.

A new Australian recording of the QuickSIN sentences was available and the purpose of this study was to test the equivalence of the 12 lists in the new recording, and to provide an SNR50 value for young Australian adults with normal hearing in order to calculate SNRloss in an equivalent manner to the original QuickSIN. The experiments were designed to test the three hypotheses that the 12 Australian lists would be of equivalent difficulty, that the SNR50 value for the normally-hearing participants would be 2 dB as for the American test, and that the benefit of HAs would be about 7 dB as measured for adaptive directional microphones in a previous study [6] using a different speech in noise test.

2. The Australian SpIN™

2.1. Recordings

Twelve SpIN lists, consisting of six sentences per list with five key words per sentence, were recorded by an Australian female speaker together with babble for 4 Australian talkers at the correct SNRs in separate channels. For hearing impaired listeners the SNR range was 20 to –5 dB and for normal hearing listeners the SNR range was 15 to –10 dB, with 5 dB decrements from one sentence to the next. The SNR ranges chosen for the Australian SpIN in order to include the possibility of negative SNR50 values as shown in previous studies [The method used to measure and equalize the levels of sentences and babble was to estimate the intensity in 20 ms windows, discard all estimates with intensity below a threshold of 30 dB and then use the 95th percentile of the remaining estimates as the intensity for the sentence or noise. The goal of this process was to ignore periods of silence before, after, or during the sentences.

2.2. Presentation

Each participant was provided with the following instructions. "Imagine that you are at a party. There will be a woman talking and several other talkers in the background. The woman's voice is easy to hear at first, because her voice is louder than the others. Repeat each sentence the woman says. The background talkers will gradually become louder, making it difficult to understand the woman's voice, but please guess and repeat as much of each sentence as possible." Each sentence in noise was presented by a computer program when the participant was ready. The test was routed through two speakers in the free field of a sound-proof booth. One speaker presented the sentences at 0˚ azimuth, with the multitalker babble noise played by a second speaker at 180˚ azimuth. The front and back locations of speech and noise were chosen to demonstrate the benefit of directional microphones to HA users. The sentences were presented at a fixed level of 65 dBA and multitalker babble was increased by 5 dB increments, to span a 25 dB SNR range. The participant repeated the words in each sentence and the experimenter indicated on the computer screen which key words were correctly repeated by the participant. One point was scored for each of the five key words correctly recognized in each sentence

3. Participants

Twenty four people participated in this study, one group of twelve young adults with normal hearing, and a second group of twelve experienced adult HA users.

3.1. Audiometric screening

All participants were tested with an audiometer to establish hearing thresholds from 250 Hz to 6 kHz. Normal middle ear function, nil communication problems, and native English speaking backgrounds were also established prior to testing.

3.2. Cognitive and memory screening

All participants were screened for cognitive deficits [10] using the Mini Mental State Examination (MMSE) [11] and the Wechsler forward and reverse digit span (DS) test [12]. A minimum score of 24 on the MMSE, a forward digit-span of 5 or more, and a reverse digit-span of 4 or more [13] were required for study participation. These criteria were required because poor auditory working memory and reduced cognition can result in poorer speech recognition performance, particularly for older adults with a hearing loss [14].

4. Standardisation with normally-hearing listeners

Twelve volunteers with hearing thresholds of ≤ 15 dB HL were recruited from the 2014 Melbourne University M. Aud. student cohort. Ages ranged from 21 to 31 years with a mean of 23.8 years. Participants listened to the 12 SpIN lists within a latin square experimental design to avoid bias effects of order, participant, and list. The SNR for the six sentences started at 15 dB and reduced by 5 dB per sentence down to –10 dB. If the actual SNR50 was 0 dB, the expected number of key words correct would be 17.5 per list. The number of key words recognized correctly was totaled for the 6 sentences of each list and the SNR50 was estimated for each list using the formula:

$$SNR50(normal\ hearing) = 17.5 - (words\ correct) \qquad (1)$$

A three-factor balanced ANOVA of the 144 SNR50 values indicated significant differences between SpIN lists {F(11,110) = 8.0, p<0.001} and participants {F(11,110) = 4.4, p<0.001} but order of list presentation {F(11,110) p=0.252} had no significant effect.

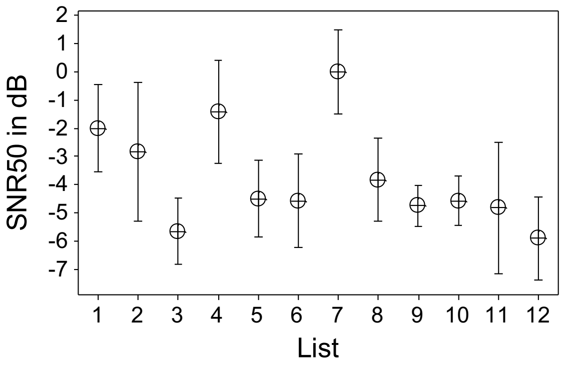


Figure 1. *Mean SNR50 scores and 95% confidence intervals for normal hearing participants (n=12) for 12 SpIN lists.*

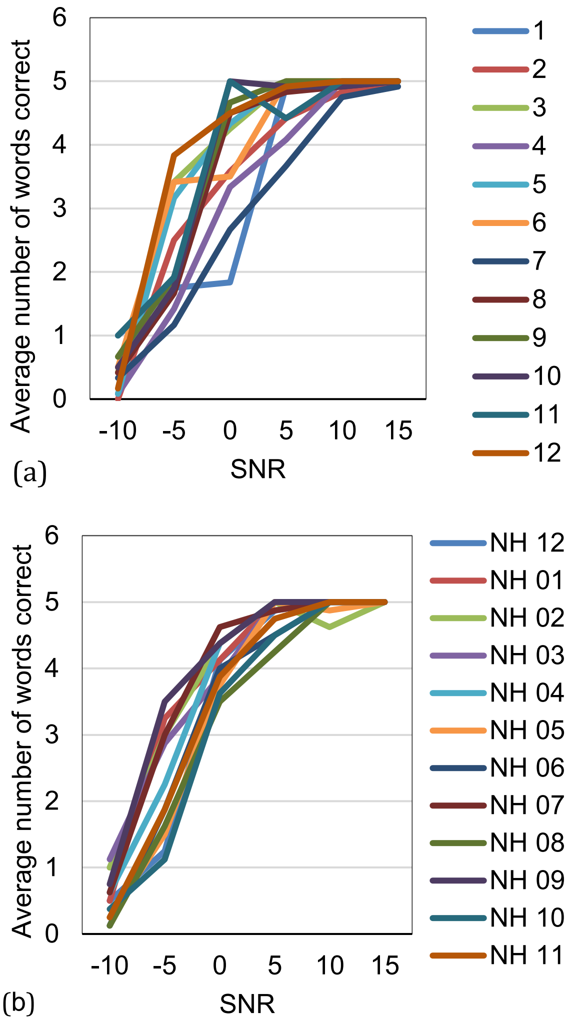


Figure 2. *Number of words correct per sentence a) averaged across participants for each list, and b) averaged across lists for each participant*

Bonferroni post hoc comparisons of the means in Figure 1 showed that lists 3, 4, 7 and 12 were different from the rest and that there was no significant difference between the remaining

8 lists. The across-subject SNR50 average for the 8 equivalent SpIN lists was –4.0 dB (± 2.7 dB st. dev.).

Figures 2 a) and 2 b) indicate that the eight balanced lists and the twelve participants produced similarly shaped performance versus SNR curves. Note that the first sentence in each list was at +15 dB SNR and every participant scored 5 words correct for every list. The last sentence in every list was at –10 dB SNR and the average number of words correct was between 0 and 1 for every list and for most participants.

5. Investigation of the SpIN with HA users

Twelve adult HA users with mild, moderate, or severe symmetrical high frequency sensorineural hearing loss were recruited from the Blamey Saunders Clinic. Their average age was 65.2 years and individual age ranged from 26 to 88 years. They were binaurally fitted with Blamey Saunders HAs which they had used for between 6 weeks and 5 years at the time the SpIN was carried out. All but one participant (11) reported using their HAs for at least 8 hours per day. The HAs all incorporated automatic adaptive directional microphone technology [6].

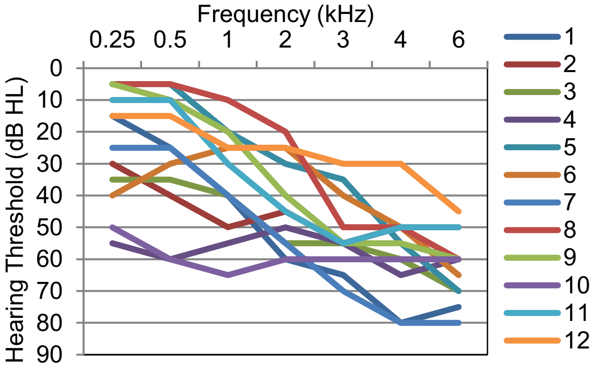


Figure 3. *Audiograms in the better ear for each participant in Group 2.*

Each participant was tested with three SpIN lists in the unaided and aided conditions. Three pairs of lists of equal difficulty were used (8 and 5, 6 and 10 and 9 and 11). One list from each of these pairs was administered in the aided and the other in the unaided condition randomly to ensure equal difficulty of sentence lists in aided and unaided conditions across participants. The condition presented first was randomly chosen. SpIN sentences in each list were presented at SNRs ranging from 20 dB to −5 dB with 5 dB decrements (ie 5 dB greater SNR than for the normally hearing participants). If the actual SNR50 was 0 dB, the expected number of key words correct would be 22.5 per list. The SNR50 for hearing impaired participants was calculated using the equation:

$$SNR50(hearing\ impaired) = 22.5 - (words\ correct) \quad (2)$$

SNR50(unaided) and SNR50(aided) were calculated by averaging over 3 lists each. SNRloss was calculated as

$$\begin{aligned} SNRloss &= SNR50(unaided) - SNR50(normal) \\ &= SNR50(unaided) + 4.0 \quad (3) \end{aligned}$$

SNRbenefit for each participant was calculated as

$$SNRbenefit = SNR50(unaided) - SNR50(aided) \quad (4)$$

The SNR50 for hearing impaired participants ranged from –3.2 to 11.2 dB in the aided condition and –0.5 to 21.2 dB in the unaided condition. SNRloss ranged from 0.8 to 15.2 dB, and SNRbenefit ranged from 0.0 to 18.3 dB.

Figure 4 shows that the shapes of the performance versus SNR curves for the HA users were less steep than the corresponding curves for the normally hearing participants in Figure 2b. Participants 1, 2, 3, 4, 7, and 10 scored 2.5 words correct or less on average in the unaided condition at 20 dB SNR. These were the participants with the greatest hearing loss in Figure 2, and their poor scores were caused by poor audibility of the sentences at the high SNR values. In this case, the formula used to estimate the unaided SNR50 value for these participants was invalid because there was no SNR tested at which the percentage of words correct was greater than 50%. The SNR50 values estimated for these participants in the unaided condition should be treated with caution. If these six participants are ignored, the unaided SNR50 for the remaining HA users ranged from – 0.2 to + 8.5 dB, SNRloss ranged from 3.8 to 12.5 dB, and SNRbenefit ranged from 0 to 9.3 dB. The mean SNR benefit was 4.1 dB. All of the HA users scored higher than 50% at 20 dB SNR in the aided condition, and so the aided SNR50 estimates are valid.

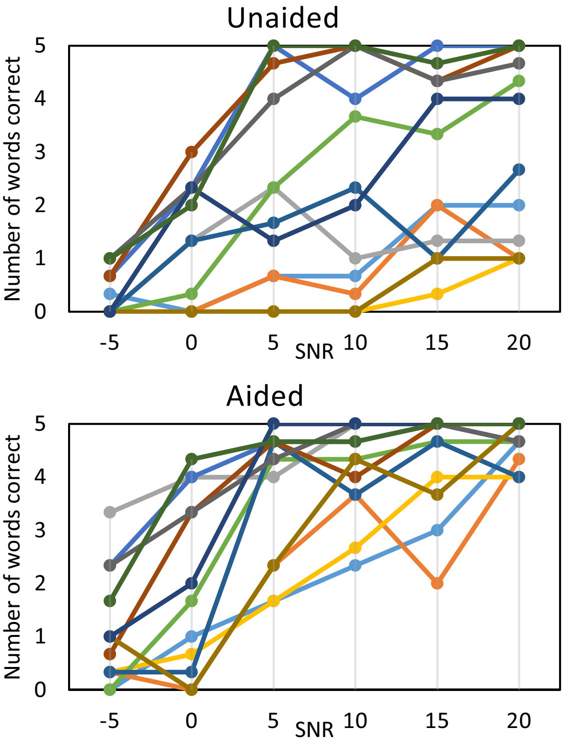


Figure 4. *Number of words correct per sentence averaged across lists for each participant in the unaided and aided conditions.*

6. Discussion

The two hypotheses based on the American QuickSIN were both disproved by this study (a) that the 12 lists would be equivalent, and (b) that the SNR50 for normal listeners would be 2 dB SNR on average.

The Australian SpIN differs from the American QuickSIN in a number of ways that had a strong influence on the results, even though the actual sentences and key words were identical.

- The sentences and babble were recorded by Australian speakers

- The sentences and noise levels were normalised in a specific manner that may have been different from the method used for the American QuickSIN
- The sentences were presented from in front of the listener (0° azimuth) and the noise was presented from behind (180° azimuth).
- Performance of normally hearing individuals was determined without filtering the sentences in Australia, and using high pass filtered speech in America.
- The Australian lists were normalised using a 15 dB to –10 dB range of SNRs while the American lists were normalised using a range from 25 dB to 0 dB SNR. In the American study, normal listeners made errors only at the 5 and 0 dB SNRs so there may have been a floor effect in their data.
- In the current study normal hearing was defined as thresholds of ≤ 15 dB across the 0.25 to 6 kHz frequency range, whilst normal hearing was defined as thresholds ≤20 dB for the same frequency range by Killion et al [4].

The mean SNR50 for normally hearing listeners was –4.0 dB in Australia and +2 dB for the American version. The differences listed above are more than sufficient to account for this. It should be noted that the Australian SpIN SNR50 is in accord with the value of –4 dB measured by Miller, Heise, and Lichten [15] for words in sentences in white noise and the value of –6.4 dB reported by Bronkhorst and Plomp [16] for sentences in speech shaped noise modulated by the envelope of 4-talker babble.

Interpretation of the unaided SNR50 estimates was complicated by the fact that half of the participants with impaired hearing did not have sufficiently good thresholds to be able to score above 50% on the sentences even when the SNR was +20 dB. When these six participants were excluded from the analysis, the data analysis was more straight-forward. Both the SNRloss values and the SNR benefits were more moderate and more in line with expectations. The mean and maximum SNR benefit of 4.1 and 9.3 dB respectively were reasonably in accord with the expected benefit of 7 dB expected from the adaptive directional microphone [6] under these testing conditions.

Although the unaided SNR50 estimates were not valid for the six HA users with the greatest hearing loss, the raw scores from the SpIN were still valid and they show that these participants were the ones who actually benefited most from their HAs in background noise. The reason for this is that there was an audibility benefit from the HAs in addition to the SNR benefit from the adaptive directional microphones.

7. Conclusion

The Australian SpIN test gave different results from the American QuickSIN from which it was derived, but in some ways, the Australian SpIN results were more in accord with the published literature on speech perception in noise. Eight of the twelve lists were chosen as being reasonably equivalent, and a study of twelve listeners with impaired hearing showed that there was a significant improvement in Australian SpIN scores between the unaided and aided conditions. Valid measurements (in dB) of SNRloss and SNRbenefit could be derived for 50% of listeners with the least degree of hearing loss. The test is useful clinically to demonstrate the benefits of HA use in background noise for clients across a wide range of hearing loss types and degrees.

8. Acknowledgements

This study was approved by the Royal Victorian Eye and Ear Hospital Human Research and Ethics Committee. The authors wish to thank the participants for their voluntary assistance in the study. Australian SpIN™ is a trademark of Blamey and Saunders Hearing Pty Ltd.

9. References

[1] Kochkin, S., "MarkeTrak VIII: Customer satisfaction with hearing aids is slowly increasing", Hear. J., 63(1):11-19, 2010.

[2] Carhart, R., and Tillman, T. W., "Interaction of competing speech signals with hearing losses", Arch. Otolaryngol., 91(3): 273-279, 1970.

[3] Nilsson, M., Soli, S. D., and Sullivan, J. A., "Development of the Hearing in Noise Test for the measurement of speech reception thresholds in quiet and in noise", J. Acoust. Soc. Am., 95(2):1085-1099, 1994.

[4] Killion, M. C., Gudmundsen, G. I., Niquette, P. A., Revit, L. J., and Banerjee, S., "Development of a quick speech-in-noise test for measuring signal-to-noise ratio loss in normal-hearing and hearing-impaired listeners", J. Acoust. Soc. Am., 116(4 Pt 1): 2395-2405, 2004.

[5] Plomp, R., "Auditory handicap of hearing impairment and the limited benefit of hearing aids", J. Acoust. Soc. Am., 63(2):533-549, 1978.

[6] Blamey, P. J., Fiket, H. J., and Steele, B. R., "Improving speech intelligibility in background noise with an adaptive directional microphone", J. Am. Acad. Audiol., 17(7):519-530, 2006.

[7] http://www.etymotic.com/auditory-research/speech-in-noise-tests/quicksin.html, "QuickSIN User Manual".

[8] www.etymotic.com/downloads/dl/file/id/.../quicksin_user_manual.pdf. "QuickSIN™ manual", Etymotic Research. Elk Grove Village IL, 2006

[9] Wilson, R. H., McArdle, R. A., and Smith, S. L., "An evaluation of the BKB-SIN, HINT, QuickSIN, and WIN materials on listeners with normal hearing and listeners with hearing loss", J. Sp. Lang. Hear. Res., 50(4):844-856, 2007.

[10] Souza, P. E., Boike, K. T., Witherell, K., and Tremblay, K., "Prediction of speech recognition from audibility in older listeners with hearing loss: Effects of age, amplification, and background noise", J. Am. Acad. Audiol., 18(1):54-65, 2007.

[11] Folstein, M. F., Folstein, S. E., and McHugh, P. R., "Mini-mental state. A practical method for grading the cognitive state of patients for the clinician", J. Psychiat. Res., 12(3):189-198, 1975.

[12] Weschsler, D., "Wechsler Adult Intelligence Scale" (4th edition). San Antonio, America: Pearson Education Inc. 2003.

[13] Choi, H.J., Lee, D.L., Seo,E.H., Min, M.K., Sohn, B.K., Choe, Y.M., Byun, M.S., Kim, J.W., Kim,S.K., Yoon, J.C., Jhoo, J.H., Kim,K.W., and Woo,J.I., "A Normative Study of the Digit Span in an Educationally Diverse Elderly Population", Psychiatry Investigation, 11:39-43, 2014.

[14] Akeroyd, M. A., "Are individual differences in speech reception related to individual differences in cognitive ability? A survey of twenty experimental studies with normal and hearing-impaired adults", Int. J. Audiol., 47(2): S53-S71, 2008.

[15] Miller, G. A., Heise, G. A., and Lichten, W., "The intelligibility of speech as a function of the context of the test materials", J. Exp. Psychol., 41(5):329, 1951.

[16] Bronkhorst, A. W., and Plomp, R., "Effect of multiple speechlike maskers on binaural speech recognition in normal and impaired hearing", J. Acoust. Soc. Am., 92(6):3132-3139, 1992.

Using Optical Flow and Electromagnetic Articulography in Multimodal Speech Research

Samantha Gordon Danner[1]*, Louis Goldstein*[1]*, Eric Vatikiotis-Bateson*[2]*, Robert Fuhrman*[2]*,*
Adriano Vilela Barbosa[3]

[1]University of Southern California, USA
[2]University of British Columbia, Canada
[3]Federal University of Minas Gerais, Brazil

sfgordon@usc.edu

Abstract

This paper describes a technique for studying the coordination of different modalities in multimodal speech using audio, video, and kinematics data. We used an electromagnetic articulograph (EMA) and audio/video recording equipment to collect data. The data is processed using EMA data analysis software and software called FlowAnalyzer, which employs the computer vision technique *optical flow* to estimate the velocity of video-recorded movement. Motion coordination between speech and manual gesture at various temporal offsets can then be analyzed using correlation map analysis and other methods. The approach outlined below represents an accessible way to study various aspects of multimodal communication.

Index Terms: Spontaneous speech, gesture, electromagnetic articulography, optical flow

1. Introduction

There is a long tradition of research on speech and co-speech gesture (henceforth *multimodal speech*) using video and audio recordings to measure coordination of acoustic features of speech with visually-observed gestures of, e.g., the hands, arms, and head [1]. These studies often report compelling results that suggest the presence of high-level coordination between speech acoustic events and body movements [2].

One potential shortcoming in many of these studies is that different kinds of measurements and temporal landmarks are used in each methodology [2]. Additionally, though there are highly developed coding methodologies for co-speech gesture research [3, 4], important details of the coding strategy – such as phases of gestural movements – can be interpreted variably, dependent on the experiment, researcher, or theory. Another critical issue is the naturally fluctuating coordination found in biological systems [5]: as exactly synchronic movements rarely occur in nature, findings of precise coordination are quite rare. Researchers may therefore wish to address the likelihood of movement correlation at temporal offsets.

The methods described here make it possible to investigate and quantify multimodal speech in novel ways. A goal of this work is to make the process of collecting and analyzing multimodal speech data more accessible to researchers. To address the concern of landmark-based measurements, we describe a method that allows for time-varying comparison of speech articulator and bodily gesture movement velocities. These velocity measurements provide an alternative to problematic coding strategies: velocity peaks – indicative of movement amplitude – in different movement signals can be compared without making assumptions about movement initiation and completion or other arbitrary landmarks in the auditory or visual domains of analysis. We also describe how correlation map analysis [5] and other tools can be used to analyze naturally fluctuating coordination at temporal offsets.

The rest of the content of this paper is organized as follows: Section 2 describes some findings and open questions in multimodal speech research. Section 3 describes our methods for studying multimodal speech coordination. Section 4 outlines some research findings and future directions for the methodology under discussion here.

2. Multimodal Speech

In experimental speech research, acoustic studies dominate the field. Acoustic data is easy to collect and analyze, but the biggest limitation is that only one modality of speech can be considered. As video recording and motion capture technology has become widely available, these tools have been integrated into speech research, but data collection and analysis of multimodal data is more complex than the analysis of acoustics alone. Studying auditory, visual *and* kinematic modes of speech together can provide a holistic view of the production and perception of speech, as there has long been an awareness that visual information [6], speech rate and amplitude modulation [7], facial expression [8] and body posture and gestures [4] are all critically important to speech perception and production.

In the domain of speech acoustics and speech articulator kinematics, some findings about temporal coordination and stereotypical gestural units [9] have become recognized as standard by researchers, but the same cannot be said for manual gesture. One issue is a lack of agreement in the research community about gesture classification schemes [10]. These schemes may disagree about how initiation and completion of gestures are defined, hand-specificity (as manual gestures may involve the use of the right hand, left hand, or both), how to handle ‘combination’ gestures (composed of two or more *types* of gesture), etc. A few studies that have taken on the task of researching multimodal speech in laboratory settings are described below.

2.1. Multimodal speech in the laboratory

Coordination of speech and manual movement has been observed in certain directed tasks in laboratory settings. Because of the inherent difficulties in the segmenting of complex time-varying co-speech gestures, many researchers have studied specific types of manual gesture in isolation, such as pointing or *deictic* gestures [4]; rhythmic finger tapping is another prevalent gestural paradigm that researchers use to stand in for more complicated gestures. Participants in such studies are typically asked to start movement of the hand or finger from, and return to, a base position to make analysis of gesture initiation and completion less ambiguous. These studies have found, across a variety of measurement paradigms, that deictic gestures and finger tapping motions are significantly correlated with positions of prosodic prominence in the speech signal [11, 12]. Such results are highly encouraging, but they suggest the need for studying a wider variety of gestures in more naturalistic speech contexts.

2.2. Open questions in multimodal speech research

Manual gesture has many more degrees of freedom than does speech articulation, and its' use is generally arbitrary. The use of manual gesture also appears to be idiosyncratic and/or culturally influenced [13]. Some researchers have questioned the underlying purpose of manual gesture, ascribing its use to, e.g., lexical retrieval rather than communication, given that speech-accompanying gesture on its own may be less communicative than speech itself [14].

The speech and gesturing tasks sometimes used in multimodal speech experimentation also raise concerns about ecological validity, especially with regard to spontaneous speech and interpersonal communication. To generate testable hypotheses regarding the coordination of speech and manual gesture, some studies investigate coordination through higher-level phenomena such as speech prosody [11]. Findings in this line of research are promising, but it is probably the case that prosody is not the only process controlling the deployment and coordination of manual gesture (respiration has been shown to play a role [15], for example). The causal relationship between speech prosody and gestural coordination is also not well understood. It is thus necessary to find a way to research coordination in a variety of communication tasks and environments.

3. Methods for studying multimodal speech coordination

The research presented here is proof of concept for the simultaneous recording of EMA kinematics data, and video data transformed into motion data with the use of optical flow software [12, 13]. Optical flow (OF) is a computer vision technique that tracks changes in pixel intensity across frames in a video sequence. OF algorithms track the magnitude and direction of these intensity changes across video frames, thus allowing for the recovery of velocity and direction of motion. FlowAnalyzer, the software that is used in this analysis, allows for post-hoc selection of regions of interest within a video prior to processing data. The software creates signals that can be additionally processed using a set of MATLAB tools [16] in subsequent steps of analysis. Minimally, a video camera with a microphone and a tripod is sufficient equipment to collect movement data with a reasonable degree of accuracy (compared with marker-based tracking [17]), thus allowing for easily portable experiment setups and research in the field. OF measurements also have the benefit of being non-invasive, so they can be used to easily track motion in regions that cannot be tracked by EMA, either due to the limited size of the equipment's magnetic field, or due to limitations on the number of sensors that can be tracked simultaneously. Body movement regions of interest measured with OF need not be predetermined, unlike with flesh point measures. The combination of EMA and video-derived OF measures allows for the direct comparison of speech articulator motion and motion of manual/bodily gesture. Previously, obtaining measures to make this comparison has been much more time-consuming and limited in scope.

3.1. Data Acquisition

The research described here investigates whether properties of gesture and coordination of speech and speech-accompanying manual gesture are task-dependent. This question was motivated by the observation that some speech situations may necessitate the use of manual gesture to a greater degree than others [18], and that different speech tasks may demand different kinds of speech-gesture coordination. The experiment is briefly described below.

3.1.1. Materials

EMA data was collected using an electromagnetic articulographer (WAVE, Northern Digital), sampled at 400 Hz. Audio, sampled at 44.1kHz, was collected concurrently using a microphone synchronized with the EMA equipment. EMA sensors were placed on the tongue body, tongue tip, lower incisor (jaw), and the upper and lower lip. Additional reference sensors were placed on the left and right mastoid processes and the upper incisor. A GoPro Hero4 video camera on a tripod positioned approximately 2' from the participant, with the speaker's head, shoulders, arms and torso in the field of view, was used to capture video at 29.97fps and audio at 48kHz.

3.1.2. Method

Participants (*N*=3) were seated comfortably in an armless chair (so as not to impede movement of the arms and hands), positioned in front of a computer monitor and next to the EMA magnetic field generator. EMA sensors were attached to the participant and the EMA equipment was calibrated. The experiment was presented in two randomized blocks. In the first block (the *demo task*), the participant was asked to demonstrate to the experimenter how they would perform an action such as eating a banana, or opening an umbrella (henceforth, these actions will be referred to as *themes*). In the second block (the *response task*), participants answered preference questions asked by the experimenter regarding each theme (e.g., "Do you prefer to eat bananas sliced or whole, and why?" or "Do you prefer umbrellas with a button closure or a Velcro closure, and why?"). Participants had a clear view of both the experimenter and the monitor. Instructions for each trial, accompanied by a black and white line drawing of the trial's theme, were presented on the monitor. The use of manual gesture was never explicitly mentioned, although the experimenter demonstrated a practice trial for each block with the use of speech and manual gesture.

3.2. Data Processing

Raw EMA data files were processed in kinematic analysis software (Mview, Haskins Laboratories), from which tangential velocities for each EMA sensor were calculated. A Praat Textgrid with interval tiers for acoustic duration, words, pauses and comments was created and annotated for each trial[19]. Praat was also used to cross-correlate audio from different sources for signal alignment.

Videos were segmented into trial-length segments and converted to .mov format at 30fps, 1280x720pixel resolution, with 44.1kHz mono channel audio (note that video need not be high resolution to work with OF). Video files were then processed in OF analysis software (e-mail the corresponding author for information on how to obtain the free FlowAnalyzer software), which creates a number of signal files that can be manipulated in MATLAB using the Audiovisual Speech Processing (AVSP) toolbox. The AVSP toolbox was used to create MATLAB structures containing the signal information obtained from the OF analysis completed in the previous step.

3.3. Data Analysis Tools

When EMA tangential velocity, acoustic segmentation and OF signals have been processed, time-varying correspondence can be analyzed using correlation map analysis (CMA) [5]. The signals generated from the AVSP toolbox and Mview can be manipulated in MATLAB or other programs to obtain measurements of interest. For example, we used MATLAB to analyze velocity time series for each movement signal (EMA sensors and OF regions of interest) in each trial. This velocity information was used to create measures of peak velocity and path length, and to measure correlations and causality in right hand and jaw velocity signals. Some findings from this analysis are described below, in section 4.

4. Findings and Future Directions

Comparison of peak velocities in the *demo* task and the *response* task (see 3.1.2) suggest that use of manual gesture may indeed be task-dependent. The peak velocities of articulator movements (Right Hand/RH and Jaw are discussed here) represent a way of measuring the average amplitude of movements without the need to predefine movement anchor points. We found that Jaw peak velocities did not differ significantly as a function of the type of speech task participants were engaged in (t(2)=0.861, p=0.4798), but average RH peak velocities were significantly greater in the *demo* task than the *response* task (t(4.92)=2.599, p=0.0491); see Figure 1. This result shows that the effect of task has a greater impact on manual gesture than it does on speech articulatory gestures, and the result was obtained without the need for a priori assumptions about manual gesture typology.

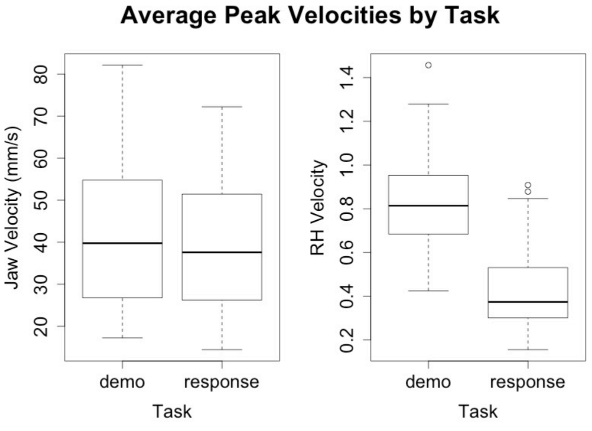


Figure 1: *Average peak velocities of movement signals in two tasks (all participants)*

We were also interested in assessing the causal relationship between two movement signals, in this case, the jaw and the right hand. We have applied the Granger test to measure causality, where a signal X Granger-causes a signal Y "if an auto-regressive model for Y in terms of past values of both X and Y is statistically significantly more accurate than that based just on the past values of Y" [20]. We applied this test to the right hand (RH) and jaw signals in each experiment condition for one theme in one speaker's data. In this small sample, the only significant result of Granger causality (see Table 1) was found in the Response condition. There, the RH signal was found to Granger-cause the Jaw signal, indicating that the instantaneous velocities of the RH signal were predictive of instantaneous velocities in the Jaw signal, at a short lag (0.115s) with respect to the RH signal. This indicates an especially close relationship between movements of the right hand and the jaw in the *response* condition, but only in one of the two possible orders. In the same condition, Jaw movements were *not* predictive of RH movements at the same short lag. The causality findings suggests that (Granger) causal relationships between speech articulator and manual movements are task-dependent, and that speech-accompanying gesture in the *response* task is more closely coordinated with speech than is gesture in the *demo* task.

Table 1: *Granger Causality tests on RH & Jaw signals in speaker M1's "fold laundry" theme*

	RH → JAW	**JAW → RH**
Demo	F(16, 4437)=0.5681, p=0.9095; lag=0.08s	F(16,4437)=0.2136, p=0.9996; lag=0.08s
Response	F(23,13095)=2.0134, p=0.0028*; lag=0.115s	F(23,13095)= 0.2974, p=0.9995; lag=0.115s

Above, we've described methods for easily obtaining measurements and comparisons of selected speech articulators and movements of the head and hands. Nearly any movement region of interest can be measured with the OF technique described here. Another benefit of this technique is that it allows for experimental designs that make use of realistic communicative events. We have also show that spoken language and manual speech-accompanying gesture each have a directly comparable kinematic component, and the relationship between the kinematic components may be used to investigate hypotheses about coordination of speech and gesture. In the case described in section 3, we have used video recording of manual and head gesture and EMA to record speech articulatory kinematics from specific points on the vocal tract. There are many other areas of study in multimodal speech research, a few of which are outlined below.

4.1. Additional scales of comparison

The OF techniques described here could be used on finely detailed movements of the face (the perioral region and the eyes and eyebrows are known to be particularly informative [8], [17]). For example, one might want to corroborate EMA recordings of lip or jaw kinematics with video recordings of the perioral region to ensure reasonable correlation of the two motion signals. Smaller regions of interest around the eyes may also be considered: As discussed in [17], potentially coordinated movement behaviors such as blinking can be captured with OF, but not with other marker-based systems.

Motion capture and OF can similarly be used for studying coordination at even larger scales. Prior research has found, for example, a relationship between postural control and vocal effort in speech [15]. This research hints at the expansive coordination of systems within the human body. Findings in this area could also help reveal what makes certain gesturing body parts like the arms, hands and head ‘special’ in multimodal speech communication.

4.2. Studying multimodal speech cross-linguistically

Many advances in the field of multimodal speech/communication have been driven by the need to investigate sign languages [21] in a principled fashion. Because (visually-observed) movements are a primary modality in sign language, the need for studying coordinated movement behavior in sign language is obvious.

Another possibility to investigate is cross-linguistic coordination and/or timing differences in various speech modalities. As EMA research has uncovered that some of the detailed timing and organization of speech articulator gestures is language-specific [22], it is reasonable to expect that comparable cross-linguistic differences exist in the timing of manual gestures with respect to speech articulator gestures and/or other bodily gestures.

4.3. Conversational Interaction

A final scenario to consider is the interaction of two or more speakers engaged in a shared speech task. The possibility of using dual EMA systems to simultaneously record vocal tract kinematics from two interacting participants has been validated [23], and the ability to video record and compute optical flow on regions of interest in conversational interactions has also been demonstrated [24]. These researchers and many others have found *entrainment* between speakers in conversation occurring at various levels. Entrainment generally refers to the rhythmic alignment that occurs between motor subsystems; the ability of the subsystems to become entrained is thought to be an indication of the presence of a higher-level dynamical system governing these subsystems. The methods described here could be applied to entrainment phenomena in multimodal speech.

5. Summary

As workflows like the one presented here become easier to implement, and as motion capture and video recording equipment becomes more affordable and accessible for researchers, we expect that the analysis of movement and kinematic data in multimodal speech research will become ubiquitous. The goal of this work is to provide some suggestions and resources for collecting and analyzing multimodal speech data, and to that end we have described research scenarios that could make use of this type of data.

6. Acknowledgments

This work was supported by NSERC and SSHRC grants to Eric Vatikiotis-Bateson and an NIH grant to Dani Byrd. The EMA data described in this paper was collected in collaboration with Sungbok Lee.

7. References

[1] S. Shattuck-Hufnagel, P. L. Ren, and E. Tauscher, “Are torso movements during speech timed with intonational phrases?,” in *Proeedings of Speech Prosody 2010*, 2010, pp. 2–5.

[2] P. Wagner, Z. Malisz, and S. Kopp, “Gesture and speech in interaction: An overview,” *Speech Commun.*, vol. 57, pp. 209–232, 2014.

[3] S. Duncan, “Annotative Practice (Under Perpetual Revision),” in *Gesture & Thought*, 2005.

[4] D. McNeill, *Hand and mind: What gestures reveal about thought*. University of Chicago press, 1992.

[5] A. Barbosa, R.-M. Déchaine, E. Vatikiotis-Bateson, and H. Yehia, “Quantifying time-varying coordination of multimodal speech signals using correlation map analysis,” *J. Acoust. Soc. Am.*, vol. 131, no. 3, p. 2162, 2012.

[6] H. McGurk and J. Macdonald, “Hearing lips and seeing voices.,” *Nature*, vol. 264, pp. 691–811, 1976.

[7] C. E. Williams and K. N. Stevens, “Emotions and speech: some acoustical correlates,” *J. Acoust. Soc. Am.*, vol. 52, no. 4, pp. 1238–1250, 1972.

[8] C. Busso and S. Narayanan, “Interplay between linguistic and affective goals in facial expression during emotional utterances,” *... Semin. Speech Prod. (ISSP 2006)*, no. 2003, pp. 549–556, 2006.

[9] L. Goldstein and M. Pouplier, “The Temporal Organization of Speech,” in *The Oxford Handbook of Language Production*, M. Goldrick, V. S. Ferreira, and M. Miozzo, Eds. Oxford University Press, 2014, pp. 1–21.

[10] J.-P. de Ruiter, “Gesture and speech Production,” 1998.

[11] B. Parrell, L. Goldstein, S. Lee, and D. Byrd, “Temporal coupling between speech and manual motor actions.,” *9th International Seminar on Speech Production*. 2011.

[12] J. Krivokapic, M. K. Tiede, and M. E. Tyrone, “A Kinematic Analysis of Prosodic Structure in Speech and Manual Gestures,” in *Proceedings of the 18th International Congress of Phonetic Sciences*, 2015.

[13] S. Kita, “Cross-cultural variation of speech-accompanying gesture: A review,” *Lang. Cogn. Process.*, vol. 24(2), no. December 2014, pp. 145–167, 2009.

[14] R. M. Krauss and U. Hadar, “The Role of Speech-Related Arm/Hand Gestures in Word Retrieval,” *Gesture, speech, sign*, pp. 93–116, 1999.

[15] R. Fuhrman, “Vocal effort and within-speaker coordination in speech production: effects on postural control,” University of British Columbia, 2014.

[16] A. V. Barbosa, H. C. Yehia, and E. Vatikiotis-Bateson, “MATLAB Toolbox for Audiovisual Speech Processing,” in *AVSP 2007*, 2007.

[17] A. V. Barbosa, H. C. Yehia, and E. Vatikiotis-Bateson, “Linguistically Valid Movement Behavior Measured Non-Invasively,” *Audit. Vis. Speech Process.*, pp. 173–177, 2008.

[18] R. B. Church, S. Kelly, and D. Holcombe, “Temporal synchrony between speech, action and gesture during language production,” *Lang. Cogn. Neurosci.*, vol. 29, no. 3, pp. 345–354, Nov. 2013.

[19] P. Boersma and D. Weenink, “Praat: doing phonetics by computer.” 2016.

[20] A. Arnold, Y. Liu, and N. Abe, “Temporal causal modeling with graphical granger methods,” *Proc. 13th ACM SIGKDD Int. Conf. Knowl. Discov. data Min. - KDD ’07*, p. 66, 2007.

[21] O. Crasborn, H. Sloetjes, E. Auer, and P. Wittenburg, “Combining video and numeric data in the analysis of sign languages within the ELAN annotation software,” *Proc. Lr. 2006 Work. Represent. Process. sign Lang.*, pp. 82–87, 2006.

[22] L. Goldstein, I. Chitoran, and E. Selkirk, “Syllable Structure as Coupled Oscillator Modes: Evidence from Georgian vs. Tashlhiyt Berber,” *Proc. XVI Int. Congr. Phonetic Sci.*, pp. 241–244, 2007.

[23] E. Vatikiotis-Bateson, A. V. Barbosa, and C. T. Best, “Articulatory coordination of two vocal tracts,” *J. Phon.*, vol. 44, pp. 167–181, May 2014.

[24] N. Latif, A. V. Barbosa, E. Vatiokiotis-Bateson, M. S. Castelhano, and K. G. Munhall, “Movement coordination during conversation,” *PLoS One*, vol. 9, no. 8, pp. 1–10, 2014.

Acoustic monitoring of speech impairment in motor neuron disease associated with frontotemporal dementia: a case series

Matthew L. Poole[1,2], *Amy Brodtmann* [2,3], *David Darby*[2,3], *Adam P. Vogel* [1,2,4]

[1]Centre for Neuroscience of Speech, The University of Melbourne, Australia
[2]Eastern Cognitive Disorders Clinic, Eastern Neurosciences, Eastern Health, Monash University, Australia
[3]Behavioural Neuroscience, Florey Institute of Neuroscience and Mental Health, Australia
[4]Department of Neurodegeneration, Hertie Institute for Clinical Brain Research, University of Tübingen, Germany

vogela@unimelb.edu.au

Abstract

Frontotemporal dementia is the second most common form of younger onset dementia. A subset of people with this disorder develop motor neuron disease (MND) with associated speech impairment (dysarthria). Here, we aim to measure the progression of dysarthria in a case of FTD-MND with acoustic analysis. Four individuals with FTD (one developing concomitant MND) were longitudinally assessed over two years. Two acoustic measures demonstrated capacity to objectively monitor dysarthria in FTD-MND. These preliminary data highlight potential for the clinical use of these methods to identify the initial signs of bulbar onset motor neuron disease.

Index terms: acoustics, disease monitoring, dysarthria, frontotemporal dementia, motor neuron disease.

1. Introduction

Frontotemporal dementia (FTD) is a form of non-Alzheimer's dementia caused by neurodegenerative atrophy in the frontal and temporal lobes of the brain [1, 2]. The behavioural variant of FTD (bvFTD) is one of three subtypes of FTD which is characterised by changes in behaviour and personality such as disinhibition, apathy, executive dysfunction and stereotypic behaviours [2]. While speech abnormalities have been noted in some people with bvFTD [3, 4], speech impairment is not regarded as a diagnostic feature of bvFTD [2].

Motor neuron disease (MND) is a neurodegenerative disorder characterised by progressive paralysis resulting from impaired functioning of the both upper and lower motor neurons [5]. The disorder can have a limb or bulbar onset, the latter affecting the muscles responsible for speech and swallowing [5]. Features of dysarthria in MND may include decreased speech rate, distortions of consonants and vowels, increased nasal resonance, reduced respiratory capacity, and a hoarse, strained voice quality [6].

While traditionally considered to be distinct disorders, MND and FTD can now be conceptualised on a clinicopathological continuum, with the pure motor involvement of MND at one end and the frontal lobe cognitive deficits of FTD at the other [7]. As many as 13-25% of MND patients meet criteria for bvFTD, while a greater proportion (up to 45%) experience some cognitive and/or behavioural changes consistent with FTD [8, 9]. Similarly, motor dysfunction which is sufficient to meet criteria for MND has been identified in 10-15% of people with FTD, while up to 36% show some signs of motor dysfunction, such as wasting, fasciculations or weakness [10]. Evidence of overlapping clinical features are supported by pathological overlap for each disorder, with the TDP-43 protein associated with a majority of familial and sporadic MND cases, and a subgroup of FTD cases [11]. Furthermore, there is a strong genetic link with the *C9orf72* intronic repeat mutation having been identified in about one third of FTD cases, half of MND cases, and the majority (70-80%) of FTD-MND cases [7, 12-14].

Assessment of bulbar involvement in MND is commonly performed by neurological examination or listener-based speech assessment, despite evidence of bulbar dysfunction occurring prior to observable changes in speech intelligibility [15]. Objective and quantifiable measures have been evaluated to identify the onset of bulbar involvement in MND and for monitoring disease progression. Acoustic correlates of bulbar involvement in this population relate to voice instability [16], temporal measures such as utterance duration and segment duration [17], and measures of the articulatory subsystem, such as vowel space area and formant trajectories [17, 18]. Rong and colleauges identified changes in articulation, in particular velocities of lip and jaw movement, and phonation (fundamental frequency at participant's highest possible pitch on a sustained vowel) which occurred prior to a discernible change in speech and intelligibility. [19] These measures are hypothesised to correlate with decreased strength of lip and jaw musculature and vocal fold weakness [19].

Accurate and early identification of disease parameters that can lead to changes in diagnosis have significant clinical and empirical value for informing disease trajectory. Data may also act as a surrogate marker for treatment response in future clinical trials. Here we present a unique longitudinal case series comparing objective speech outcomes in MND and bvFTD.

2. Method

2.1. Participants

Four participants with a diagnosis of bvFTD were recruited from the Eastern Cognitive Disorders Clinic, Melbourne (see Table 1 for demographic information). One of these cases, VP, presented with concomitant dysarthria associated with MND.

All participants were assessed at two time points, two years apart.

2.1.1. VP

At his initial assessment, VP was a 71 year old man, who had been referred to a tertiary assessment clinic for investigation of his suspected frontotemporal dementia and dysarthria. He had been symptomatic for approximately four years. He had changes to his behaviour of four years duration and signs of dysarthria in the 12 months prior to the first assessment. His wife also described gait and balance changes at the first assessment.VP satisfied criteria for bvFTD and was diagnosed as such by a neurologist with expertise in behavioural neurology. The presence of dysarthria suggested potential MND. Kennedy's disease was raised as a possible differential diagnosis. Genetic testing for Kennedy's disease was offered but declined by VP's family.

2.1.2. bvFTD case 1

bvFTD case 1 is a male participant who underwent his first speech assessment at age 59. He had a four year history of personality and behavioural change. Language, motor speech and visuospatial function were intact.

2.1.3 bvFTD case 2

bvFTD case 2 is a male participant who underwent his first speech assessment at age 70. He had received a diagnosis of behavioural variant frontotemporal dementia at age 65, following a 4 – 6 year history of behavioural and personality change. His speech was fluent and grammatical, however hesitant due to word finding difficulties.

2.1.4 bvFTD case 3

bvFTD case 3 underwent his first speech assessment at age 76, with a 15 year history of slowly progressing bvFTD. A neuropsychological assessment conducted at age 76 revealed impairments of executive functioning, naming, word comprehension, semantic knowledge, and memory retrieval.

Table 1 *Participant demographics*

	VP	bvFTD Case 1	bvFTD Case 2	bvFTD Case 3
Gender	M	M	M	M
Age time 1	71	59	70	76
Age at onset	67	54	59 – 61	61

2.2. Speech sample recording and stimuli

Participants provided four speech samples: (i) a one minute monologue about something that they enjoyed; (ii) saying the days of the week; (iii) producing a sustained /a/ vowel on one breath (maximum phonation time); (iv) repetition of multisyllabic words; (v) diadochokinetic rate (DDK; saying "pataka" repeatedly as quickly and clearly as possible). All tasks were produced twice, with the exception of the monologue, and the second iteration was used in all analyses in order to mitigate the effect of unfamiliarity on the novel tasks [20]. The tasks have been shown to have reliability and sensitivity to impairment for measures of speech timing [20, 21]. The speech samples were recorded using a Marantz PMD671 solid state recorder with an AKG C520 condenser cardioid head mounted microphone positioned 8 cm from the participants' mouth at a 45° angle. Recordings were sampled at 44.1 KHz and quantized at 8 bits.

2.3. Speech analysis

Speech was quantified objectively using acoustic analysis and subjectively via listener based evaluations at both first and second time points. Details of these methods and the stimuli utilised are outlined below.

2.3.1. Acoustic analysis of speech

Measures of speech timing, vowel articulation (the vowel articulation index), and voice (harmonics to noise ratio) were conducted. Speech timing measures (syllables per second, mean pause length (MPL), proportion of pause time (PPT)) were calculated for the days of the week stimulus using automated scripts, derived from the methodology of [22] in Praat [23]. The days of the week stimulus was selected for timing measurement as it was hypothesised to be less influenced by cognitive and behavioural impairment compared to the monologue, due to its automaticity. The vowel articulation index (VAI) was calculated by measuring the first two formants (prominent resonant frequencies) of the /a/, /i/ and /u/ vowels [24]. The /a/ and /i/ vowels were taken from repetition of the word 'artillery', and the /u/ vowel from 'Tuesday' in the days of the week task. First (F1) and second formant (F2) frequencies were calculated in Praat, and the VAI was calculated with the following formula: VAI = (F2/i/ + F1/a/)/(F1/i/ + F1/u/ + F2/u/ + F2/a/). Decreasing VAI values indicate centralisation of the vowel formants, which is indicative of impaired vowel articulation [25]. The harmonics to noise ratio (HNR) quantifies the amount of additive noise in the voice signal relative to the harmonic component [26] to provide an objective evaluation of the degree of hoarseness in a person's voice.

2.3.2. Listener-based speech assessment

The participants' speech was rated by two speech pathologists (MLP & APV). The raters were blinded to participant, diagnosis and time point. Speech samples were rated independently by each rater and disagreement was resolved by consensus. Samples were assessed on a range of speech domains with a five point severity rating scale (0 = no impairment, 1 = sub-clinical, 2 = mild, 3 = moderate, 4 = severe impairment). Twenty-five speech features were assessed for severity within the domains of pitch, respiration, loudness, prosody, voice, articulation, resonance and DDK production.

3. Results

3.1. Acoustic analysis of speech

3.1.1. Speech timing

The degree of change in participants' MPL at time points 1 and 2 is presented in Figure 1. VP presented with an MPL at the first time point of 0.058 and 0.323 at the second. This change exceeded two standard deviations of the bvFTD group mean change (0.036 at time 1, and 0.063 at time 2). Figure 2 shows the magnitude of change for proportion of silence time in each sample. At the second time point, the mean decrease for the bvFTD group was 15% (from 10.28 to 8.68). VP's proportion of silence increased by 22% (from 27.26 to 33.13), which was greater than two standard deviations of the bvFTD group mean change. Degree of change in speech rate is presented in Figure 3. The rate of change for VP was similar to that of the bvFTD group.

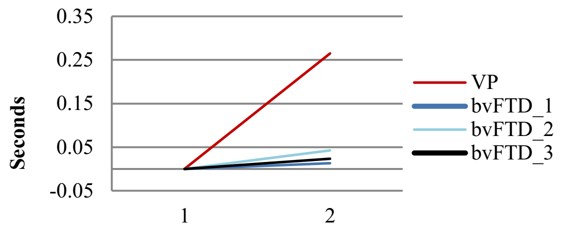


Figure 1: *Degree of change for MPL between two time points*

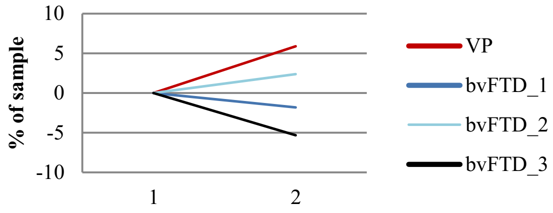


Figure 2: *Degree of change for PPT between two time points*

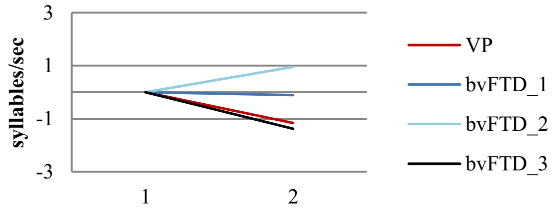


Figure 3: *Degree of change for speech rate between two time points*

3.1.2. Vowel articulation index

VAI values for each participant are presented in Figure 4. On average, the bvFTD participants' VAI changed by 1% between the two time points (0.734 to 0.733). VP showed a 15% decrease (0.861 to 0.730) in VAI which indicates a reduced qualitative distinction between the /i/ ("ee"), /u/ ("oo"), and /a/ ("ah") vowels. This change was greater than two standard deviations of the bvFTD participants' mean change.

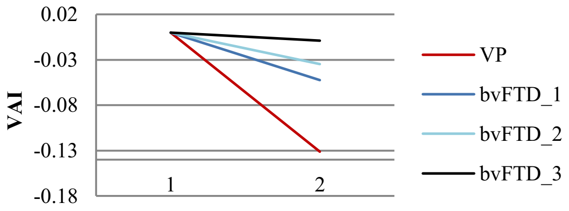


Figure 4: *Degree of change for VAI between two time points*

3.1.3. Harmonics to noise ratio

Participants' harmonics to noise ratios are presented in Figure 5. There was a high degree of variance within the bvFTD group for this measure, and VP was within one standard deviation of the control group mean change.

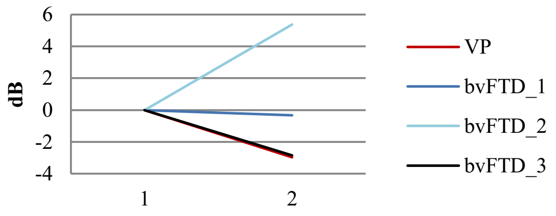


Figure 5: *Degree of change for HNR between two time points*

3.2. Listener-based speech assessment

VP had impairments of prosody, articulation and voice at the first assessment. His second assessment revealed deterioration of prosody, articulation and pitch variation. There was minimal change in perceptual ratings for the bvFTD participants.

4. Discussion

The use of acoustic measures allowed accurate detection of the accelerated change caused by the motor speech impairments of MND as opposed to the cognitive-linguistic and behavioural impairments of bvFTD. Consensus ratings from two speech pathologists blinded to participant and time point confirmed a perceptual change to VP's speech across multiple speech sub-systems, predominantly those of pitch, articulation and prosody. Deterioration of speech quality was also observed in the bvFTD participants; however, these were restricted to changes to voice quality and prosody. The study is limited as participants are recorded only at two timepoints, and further longitudinal follow up would enhance the study by demonstrating consistent change over time. These features will be discussed in relation to the acoustic correlates of timing, vowel articulation and voice quality.

4.1.1. Timing

Mean pause duration was the most effective timing method for identifying changes of speech secondary to motor dysfunction, as demonstrated by the large increase in VP's MPL relative to the bvFTD participants.

The increase in VP's MPL is consistent with the change in his speech rate (syllables per second), for which VP was shown to have a reduction of approximately one syllable per second. Two bvFTD participants experienced changes of similar magnitude, which suggests that this measure is also sensitive to behavioural and cognitive change.

Investigation of these timing measures in MND and FTD at a single time point has shown that the PPT has been shown to differentiate MND patients with predominantly respiratory symptoms [27] from bulbar onset MND, bvFTD and progressive nonfluent aphasia, and may act as a measure of respiratory function deterioration [19]. The significant increase in VP's MPL in this study may indicate that this measure has utility in measuring within-individual change.

4.1.2. Vowel articulation

Deterioration of articulation was quantified with the vowel articulation index. VAI has been shown to be effective in identifying and monitoring dysarthria in Parkinson's disease [25, 28], even prior to perceptual identification of vowel distortions [29]. Smaller vowel space areas, as demonstrated for VP, have been previously documented as a feature of dysarthria in MND in comparison to healthy controls [30, 31]. These findings suggest that vowel articulation index may be valuable for identifying and measuring articulatory change in the FTD-MND continuum.

4.1.3. Voice quality

There was variation in harmonics to noise ratio (HNR) for both VP and the bvFTD participants, with bvFTD participants experiencing both increases and decreases in HNR. Deterioration of HNR is expected in healthy ageing [32], and this is consistent with known changes to vocal fold physiology in older adults [33]. In this case series, HNR declined with increased breathiness over time, however it was not useful in identifying change that was specific to motor dysfunction as opposed to changes associated with ageing.

5. Conclusions

Several quantitative speech measures have demonstrated capacity to monitor change in both MND and bvFTD. Measures of mean pause rate and VAI were shown to be sensitive to the greater magnitude of change associated with motor speech dysfunction, in comparison to more general behavioural and cognitive changes related to FTD. These quantitative measures were largely consistent with listener-based ratings of speech changes, and therefore have potential for objective monitoring which could be utilised as an adjunct to listener-based ratings in clinical settings. In particular, such measures have potential to assist in the early identification of bulbar onset MND in the FTD population. Future longitudinal studies with larger cohorts could allow for the sensitivity and specificity of these measures to be established. A focus on assessing speech changes over shorter time periods would further clarify their potential as clinical measures.

6. References

[1] Neary, D., et al., *Frontotemporal lobar degeneration: A consensus on clinical diagnostic criteria.* Neurology, 1998. **51**(6): p. 1546-1554.

[2] Rascovsky, K., et al., *Sensitivity of revised diagnostic criteria for the behavioural variant of frontotemporal dementia.* Brain, 2011. **134**(9): p. 2456-2477.

[3] Mendez, M.F., et al., *Clinicopathologic differences among patients with behavioral variant frontotemporal dementia.* Neurology, 2013. **80**(6): p. 561-8.

[4] Diehl, J. and A. Kurz, *Frontotemporal dementia: patient characteristics, cognition, and behaviour.* International Journal of Geriatric Psychiatry, 2002. **17**(10): p. 914-8.

[5] Kiernan, M., et al., *Amyotrophic lateral sclerosis.* Lancet (London, England), 2011. **377**(9769): p. 942-955.

[6] Tomik, B. and R.J. Guiloff, *Dysarthria in amyotrophic lateral sclerosis: A review.* Amyotrophic Lateral Sclerosis, 2010. **11**(1-2): p. 4-15.

[7] Devenney, E., et al., *Motor neuron disease-frontotemporal dementia: a clinical continuum.* Expert review of neurotherapeutics, 2015. **15**(5): p. 509-522.

[8] Lillo, P., et al., *Amyotrophic lateral sclerosis and frontotemporal dementia: a behavioural and cognitive continuum.* Amyotrophic Lateral Sclerosis, 2012. **13**(1): p. 102-109.

[9] Phukan, J., et al., *The syndrome of cognitive impairment in amyotrophic lateral sclerosis: a population-based study.* Journal of Neurology, Neurosurgery & Psychiatry, 2012. **83**(1): p.102-8.

[10] Burrell, J.R., et al., *Motor neuron dysfunction in frontotemporal dementia.* Brain, 2011. **134**(9): p. 2582-2594.

[11] Neumann, M., et al., *Ubiquitinated TDP-43 in frontotemporal lobar degeneration and amyotrophic lateral sclerosis.* Science, 2006. **314**(5796): p. 130-133.

[12] DeJesus-Hernandez, M., et al., *Expanded GGGGCC hexanucleotide repeat in noncoding region of C9ORF72 causes chromosome 9p-linked FTD and ALS.* Neuron, 2011. **72**(2): p. 245-256.

[13] Renton, A.E., et al., *A hexanucleotide repeat expansion in C9ORF72 is the cause of chromosome 9p21-linked ALS-FTD.* Neuron, 2011. **72**(2): p. 257-268.

[14] Rohrer, J.D., et al., *C9orf72 expansions in frontotemporal dementia and amyotrophic lateral sclerosis.* The Lancet Neurology, 2015. **14**(3): p. 291-301.

[15] Green, J.R., et al., *Bulbar and speech motor assessment in ALS: Challenges and future directions.* Amyotrophic Lateral Sclerosis and Frontotemporal Degeneration, 2013. **14**(7-8): p. 494-500.

[16] Ramig, L.O., et al., *Acoustic Analysis of Voice in Amyotrophic Lateral SclerosisA Longitudinal Case Study.* Journal of Speech and Hearing Disorders, 1990. **55**(1): p. 2-14.

[17] Weismer, G., et al., *Acoustic and intelligibility characteristics of sentence production in neurogenic speech disorders.* Folia Phoniatrica et Logopaedica, 2001. **53**(1): p. 1-18.

[18] Weismer, G., et al., *Formant trajectory characteristics of males with amyotrophic lateral sclerosis.* The Journal of the Acoustical Society of America, 1992. **91**(2): p. 1085-1098.

[19] Rong, P., et al., *Predicting early bulbar decline in amyotrophic lateral sclerosis: A speech subsystem approach.* Behavioural neurology, 2015. **2015**.

[20] Vogel, A.P. and P. Maruff, *Monitoring change requires a rethink of assessment practices in voice and speech.* Logoped Phoniatr Vocol, 2013.

[21] Vogel, A.P., et al., *Reliability, stability, and sensitivity to change and impairment in acoustic measures of timing and frequency.* J Voice, 2011. **25**(2): p. 137-49.

[22] Vogel, A.P., J. Fletcher, and P. Maruff, *Acoustic analysis of the effects of sustained wakefulness on speech.* The Journal of the Acoustical Society of America, 2010. **128**(6): p. 3747.

[23] Boersma, P. and D. Weenink, *Praat, a system for doing phonetics by computer.* 2001.

[24] Roy, N., et al., *Articulatory changes in muscle tension dysphonia: evidence of vowel space expansion following manual circumlaryngeal therapy.* Journal of communication disorders, 2009. **42**(2): p. 124-135.

[25] Skodda, S., W. Visser, and U. Schlegel, *Vowel articulation in Parkinson's disease.* Journal of Voice, 2011. **25**(4): p. 467-472.

[26] Yumoto, E., W.J. Gould, and T. Baer, *Harmonics-to-noise ratio as an index of the degree of hoarseness.* The Journal of the Acoustical Society of America, 1982. **71**(6): p. 1544-1550.

[27] Yunusova, Y., et al., *Profiling Speech and Pausing in Amyotrophic Lateral Sclerosis (ALS) and Frontotemporal Dementia (FTD).* PloS one, 2016. **11**(1).

[28] Skodda, S., W. Grönheit, and U. Schlegel, *Impairment of vowel articulation as a possible marker of disease progression in Parkinson's disease.* PloS one, 2012. **7**(2): p. e32132.

[29] Rusz, J., et al., *Imprecise vowel articulation as a potential early marker of Parkinson's disease: Effect of speaking task.* The Journal of the Acoustical Society of America, 2013. **134**(3): p. 2171-2181.

[30] Turner, G.S., K. Tjaden, and G. Weismer, *The influence of speaking rate on vowel space and speech intelligibility for individuals with amyotrophic lateral sclerosis.* Journal of Speech, Language, and Hearing Research,1995.**38**(5):p.1001-13.

[31] Turner, G.S. and K. Tjaden, *Acoustic differences between content and function words in amyotrophic lateral sclerosis.* Journal of Speech, Language, and Hearing Research, 2000. **43**(3): p. 769-781.

[32] Ferrand, C.T., *Harmonics-to-noise ratio: an index of vocal aging.* Journal of voice, 2002. **16**(4): p. 480-487.

[33] Pontes, P., A. Brasolotto, and M. Behlau, *Glottic characteristics and voice complaint in the elderly.* Journal of Voice, 2005. **19**(1): p. 84-94.

Similarity in global accent promotes generalized learning of accent markers

Ann-Kathrin Grohe, Margarita Downing, Andrea Weber

University of Tübingen

ann-kathrin.grohe@uni-tuebingen.de, margarita.pelz@student.uni-tuebingen.de, andrea.weber@uni-tuebingen.de

Abstract

Is it easier to adapt to an accent when the speaker has an accent similar to oneself than when not? This question was addressed in a training-test study. During training, English L1 participants either listened to a story read by an L1 speaker who replaced all /θ/s with /t/, or they produced the story with the /θ/-substitutions themselves. Learning, i.e., faster lexical decision times to accented words, was observed after both production and listening training. Speaker-listener similarity was ensured by having L1 speakers during training and test. Without this similarity, no learning had been observed in [8].

Index Terms: native accent, foreign accent, adaptation, accent similarity, accent strength, auditory word recognition

1. Introduction

Native as well as foreign accents are pervasive in everyday speech. While both accent types deviate from the standard pronunciation of a language, they are produced by speakers with different native language backgrounds—foreign accents are produced by second language (L2) speakers; native accents are produced by native (L1) speakers of the target language. Both accent types can initially slow down comprehension, but L1 listeners are able to adapt to them. In [1], English L1 listeners overcame initial processing difficulties for sentences spoken by Spanish- and Chinese-accented speakers within one minute of listening to these speakers. In [2], listening to a 4-minute story in German-accented Dutch was sufficient for Dutch L1 participants to subsequently show facilitatory priming for words with a strong German accent marker (/œy/ as in *huis* 'house' pronounced as /ɔɪ/); see [3] for comparable results). Adaptation to native accents has been investigated, for example, in [4]. Students originally from Northern England adapted to Standard Southern British English within two years of their university studies in Southern England. Adaptation to native accents is also possible with a training phase that only lasts several minutes and can even affect cohort activation of unaccented words, as found in a recent eye-tracking study [5]. Adaptation was, however, speaker-specific, i.e., it was restricted to the training speaker and did not generalize readily to new speakers.

Thus, adaptation to foreign- and native-accented speech occurs, but it is not clear yet if the underlying processes are the same. One possibility is that adaptation follows the exact same principles in both cases and only the accent's acoustic distance from standard speech, i.e., accent strength, is the determining factor for adaptation ease. Support for the role of accent strength comes, for example, from [2] who found that priming effects were smaller for strongly-accented words than for weakly-accented words. On the other hand, [6] observed that native English listeners generalize learning of a position-specific accent marker (devoicing of final /d/) to a new position only when the accent marker was learned from an L2 speaker, and not when the same accent marker was learned from an English L1 speaker. This suggests more tolerant adaptation to foreign-accented speech than to native-accented speech. Further support for this notion was found in [7]. Dutch L1 listeners adapted to the same German-accented words more quickly when the speaker was perceived to be an L2 speaker of Dutch in the filler items than when they were perceived to be a native speaker of Dutch. Thus, not only accent strength, but possibly also the nativeness of the speaker, can influence the adaptation process.

In all of the above studies, adaptation occurred through listening. However, producing an accent can also form the basis for adaptation. This was tested recently in [8]. In an exposure-test paradigm, the effect of an individual's own accent production was compared to that of listening to someone produce an accent. Participants first either listened to an English short story recorded by a German learner of English who replaced all dental fricatives (*th*s) with /t/ (e.g., *theft* became **teft*), or they read the same story aloud themselves with the instruction to substitute all *th*s with /t/. Neither German learners of English [9] nor the tested population of English L1 speakers typically replace *th* with /t/, but they have no difficulties producing /t/s when instructed to do so. After the training, participants completed a lexical decision task on words with *th*-substitutions spoken by another German learner of English. Surprisingly, English L1 participants showed no adaptation, while German L2 participants accepted, both after production training and after listening training, accented words more quickly than a control group. One possible explanation for the lack of observable adaptation effects for L1 participants is based on the fact that, in addition to testing generalization of learning across two speakers, L1 participants did not share a language background with the L2 test speaker. Possibly, this combination resulted in acoustic differences between training and test items that were too large for learning to generalize across two speakers. Support for this explanation comes from a recent eye-tracking study [10] in which German L1 participants adapted to an accent marker (devoicing of word-initial voiced stops; e.g., *Balken* 'beam' became **Palken*) both after having listened to a German L1 speaker produce the accent and after having produced the accent themselves. If this pattern of results transfers to English L1 participants and to a lexical decision task with reaction times as dependent measure, then English L1 participants should adapt to an accent marker well enough to generalize effects to a new speaker using the materials in [8], but only if the materials are produced by English L1 speakers.

In summary, prior research has shown that adaptation to foreign and native accents is possible, with both listening and production training. It is not clear, though, what is the role of a speaker's nativeness in accent adaptation.

This was investigated in the present study. English L1 participants performed the same production and listening training and subsequent word recognition task as in [8]. That is, they either listened to an English story in which all *th*s were produced as /t/, or they produced the story with the accent themselves before responding to English words with the accent marker in a lexical decision task. However, both the training story for the listening group and the test items were recorded this time by two female native speakers of English rather than by German learners of English. Thus, the accent marker was the same as in [8], but in both training conditions the language background of the training speaker (L1 participants from the production training and the pre-recorded L1 speakers from the listening training) now matched the language background of the test speaker.

The present study thus tested whether the similarity between one's own accent and the accent to be learned facilitates learning. This would be in line with the assumption that L2 participants in [8] generalized accent learning across speakers because their own accent was similar enough to the pre-recorded speaker's accent. It was expected that English L1 participants in the present study would adapt to the accent and generalize learning to a new speaker both after listening and after production training because the language background of the speaker during test was the same as that of the participants, thereby increasing the global similarity of the speech stimuli. Such a result would show that in an individual's L1, speaker-listener similarity in terms of global accent can promote speaker-general learning of a specific accent marker.

2. Experiment

2.1. Participants

Fifty-nine female students (18-26 years old) from the University of Maryland, all L1 American English speakers, were tested.

2.2. Material

The training text was 565 words long and was based on the fairy tale "King Thrushbeard". Each *th* in the text was a digraph and corresponded to the English interdental fricative /θ/ in 39 words. For the lexical decision test, 24 English words with /θ/ in initial position were chosen as critical words (mean frequency: 163.6 occurrences per million according to the CELEX word form dictionary [11]. When /θ/ was replaced with /t/, the resulting word forms were nonwords (*theft* became **teft*). Twelve of the critical words were taken from the exposure story (old words), and 12 were new. Old and new critical words were matched for frequency and number of syllables. New critical words were included in the study to test if adaptation to the accent can generalize across the lexicon or is specific to the trained old words. An additional 120 filler words (24 contained a /t/, none contained a /θ/) were selected, half of them words and half nonwords.

The selected material was the same as in [8], but the stimuli were recorded by English L1 speakers rather than by German learners of English as in [8]. Training and test materials were recorded with a professional recording device (Focusrite Scarlett 2i2 and a Rhode NT1-Kit; 44.1 kHz, 16 bit) by two L1 American English female speakers (speaker A: 29 years old, from New Jersey, USA; speaker B: 35 years old, from Georgia, USA). Two speakers were recorded so that different voices could be used for training and test in the listening training group, which accounts for a comparable difference in voices in the production training group. The speakers did not differ significantly in F0-range or speaking rate. They were instructed to pronounce all *th*s as a /t/, but, otherwise, to speak as naturally as possible. Every *th*-instance was highlighted in yellow in the text for the recording. The final story of both recorded speakers was about three minutes in length with no significant difference between speaker A and B.

Table 1. *Critical old words from the training story and critical new words.*

old words	new words
thankful, theft, therapist, thing, think, thinner, thirsty, thrifty, throughout, throw, thumb, Thursday	Thanksgiving, thematic, theory, thesis, thickness, thief, thirty, threaten, threshold, thriller, throat, thunder

2.3. Design and Procedure

Twenty participants listened to the recordings of the training story (listening group), 20 read the story out loud and deliberately substituted all *th*s with /t/ (production group), and 19 had no training (control group). Half of the participants heard speaker A during test, and half heard speaker B. For the listening group, the speaker of the training story was always different from the speaker during test. Four experimental lists (including all 144 items; 24 critical words and 120 fillers) were created for the lexical decision task with varying, pseudo-randomized item order. Each critical word was preceded and followed by at least one filler. The lists were distributed equally across participants. The experiment was programmed with the software *Presentation*® (Version 18.3, www.neurobs.com).

The experiment took place in a soundproof room at the University of Maryland. Each participant was seated in front of a computer screen and wore noise-canceling headphones. All instructions were provided in written English. The listening group listened twice to the pre-recorded story with *th*-substitutions while seeing the story on the screen; the production group read the story twice out loud with the *th*-substitutions. The listening group was asked to pay specific attention to the pronunciation of the talker and to report oddities afterwards. This ensured that they were just as attentive and conscious of the substitutions as the production group naturally was. The production group followed the substitution instructions quite consistently (mean error rate of all voiceless *th*-occurrences: 2.9%).

After the training, English instructions for the lexical decision task were presented on the screen. Participants were told to decide as quickly and reliably as possible whether a presented auditory stimulus was an existing English word or not. Each stimulus was preceded by a fixation cross on the screen for 500 ms. As in Experiments 2 and 3 of [8], the participants received explicit instructions to accept critical words (e.g., **teft*) during lexical decision in order to clarify any uncertainty about the decision. The control group was given one example of the accent and also instructed to treat items with that accent as words. Explicit instructions were necessary because word forms with *th*-substitutions were not real words in English (e.g., **teft* for *theft*), but participants might consider them words after accent training. No feedback was given during lexical decision. After the lexical decision task was completed, participants filled out a language background questionnaire.

2.4. Results

Analyses were conducted with the software R (version 3.2.4, www.r-project.org). Endorsement rates and reaction times for accepted critical words (i.e., yes responses to **teft*) were analyzed with linear mixed effects models [12]. Reaction times, measured from word offset, between 85-1900 ms long were included in the analyses (5% outliers) and were log-normalized. For each analysis, we built an individual, best fitting model that included only significant fixed factors as well as random factors (participant and item as random intercepts). This was done with a backward stepwise selection procedure starting with the most complex model including all possible main effects and interactions that still converged. Significance of level comparisons was indicated by *t*-/*z*-values>|2|. Corresponding *p*-values of factors and interactions, as reported in the text below, were determined with likelihood ratio tests using the anova()-function.

2.4.1. *Endorsement Rates*

Endorsement rates were on average 86% (listening group 86.1%; production group 91.6%; control group 81.5%). Mean endorsement rates of individual items ranged between 63.6% (*thematic*) and 100% (*thankful, Thanksgiving, thirsty, threshold*). Statistical analyses show that the effect of training group was significant (χ^2=9.3, p<.01). The production group accepted significantly more accented tokens than the control group (*β*=1.1, *SE*=0.34, *z*=3.3). The explicit instruction to accept all accented tokens as words, however, renders endorsement rates less informative by making the choice to accept critical words less spontaneously. As in [8], focus will therefore be placed on reaction time analyses, which are particularly suitable to provide information about online processing.

2.4.2. *Reaction Times*

Reaction times for accepted critical tokens (i.e., yes responses to **teft*) were analyzed with a model having as fixed factors an interaction between training (with the levels production, listening, and no training) and familiarity (with the levels old words and new words), as well as item duration and list position. Item and participant were cross-random factors, and by-participant random slopes for list position as well as by-item random slopes for list position and speaker (with the levels speaker A and speaker B) were included. The training*familiarity interaction was significant (χ^2=10.9, p<.005). This interaction stems from differences between both training groups and the control group (a main effect of training) for old items. These differences were less strongly pronounced for new items. The listening group accepted old items faster than the control group (*β*=-0.31, *SE*=0.09, *t*=-3.4), and there was a strong trend to accept new items more quickly in the listening group than in the control group as well (*β*=-0.16, *SE*=0.09, *t*=-1.8). The production group tended to be faster than the control group in accepting old items (*β*=-0.16, *SE*=0.09, *t*=-1.7), but there was no effect for new items (*t*<1).

In order to further ensure that the observed effects indeed reflect a manipulation of the training conditions rather than more random differences in general processing speed between participant groups, reaction times to correctly accepted word fillers (60-1500 ms) were analyzed as a baseline comparison. Word fillers were canonical word forms without specific accent markers, and reaction times of the listening group were on average 14% faster than those of the control group, while the production group was 2.4% slower than the control group. Statistical analyses showed a significant main effect of training group (χ^2=12.2, p<.04). To correct for this bias, reaction times for accepted critical tokens in the training groups were adjusted to the processing speed of the control group (i.e., reaction times of the listening group were increased by 14% and those of the production group reduced by 2.4%).

With the new, adjusted reaction times, the training*familiarity interaction was still significant (χ^2=10.9, p<.005). Training effects were observed for old, but not new items (see Figure 1). The listening group accepted old items faster than the control group (*β*=-0.17, *SE*=0.09, *t*=-1.9), and the production group was significantly faster than the control group (*β*=-0.18, *SE*=0.09, *t*=-2). Reaction times for new items were not affected by accent training (all *t*-values<|0.5|). Next, we analyzed our data together with the L1 data from [8]. The same material and design was used in the latter study, but the pre-recorded tokens were spoken by two female L2 learners of English. However, no training effects were observed with L1 participants in [8]. The new, large dataset consequently included the additional variable *speaker nativeness* (L1 vs. L2 English). The same model as above was run, replacing the two-way interaction with the three-way interaction of training*familiarity*speaker nativeness. Adding speaker nativeness significantly improved the model (χ^2=12.2, p<.04). This confirms that speaker nativeness was the critical factor that provoked the training effects.

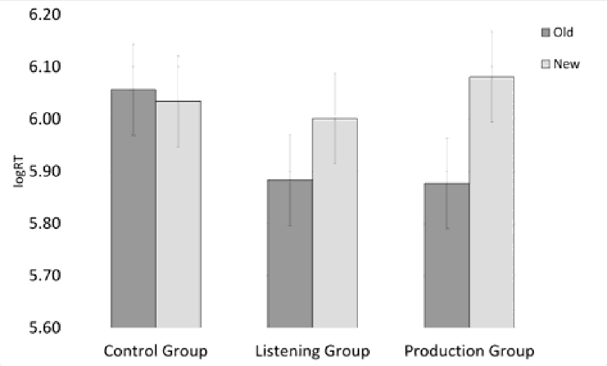


Figure 1. *Adjusted Reaction Times (with standard errors) to old and new items in the Listening Group, the Production Group, and the Control Group without training.*

3. Discussion

In the present study, L1 English participants learned an accent marker in their L1 well enough to generalize it to a new speaker through brief listening and production training. The listening training and test material was recorded by L1 English speakers. The same material, though recorded with a second language speaker, had previously not induced speaker general learning in either training condition [8]. Joint analyses of the present data and the analogous L1 experiment in [8] confirmed that accent adaptation depends in this case on the pre-recorded speaker's native language background. This underlines the importance of similar accent properties between test speakers and participants.

Accent similarity was created by having a test speaker with the same native language background (American English) as the participants. L1 participants adapted to the accent with the L1 test speaker. In addition to greater speaker-listener similarity, L1 speakers also had a smaller degree of overall accent strength. In contrast to the L2 speakers in [8], they did not have a global L2 accent. Thus, not only speaker-listener similarity but also generally weaker accentedness could have driven the present findings. The role of similarity, however, is

Table of Contents for this manuscript

strengthened by the fact that L2 participants in [8] did generalize learning of the accent when it was produced by L2 speakers with a global L2 accent.

The present findings, moreover, emphasize the role of global accent markers. Even if a global accent does not inhibit processing as much as a specific accent marker might ([3], Experiment 1), it can still play an important role for generalization of accent learning across speakers. Adding a global accent to a specific accent marker can increase or reduce speaker-listener similarity and thereby affect generalized accent learning. The reason why accent similarity between speaker and listener is so important in accent learning likely lies in participants' prior experience with the accent in question. A language user is more experienced in both producing and listening to (also by self-listening) their own accent than other accents, which facilitates accent learning.

In contrast to many prior studies on accent learning, the present study tested accent learning across speakers, i.e., whether an accent can be learned from one speaker and be applied to a second speaker with the same native language background. We found that generalization across speakers is possible as long as both speakers have similar accent properties. Adaptation across speakers was also found with L2 participants in [8], suggesting that accent adaptation is not necessarily speaker-specific as was suggested in [5]. This is in line with further studies [13–15] that suggest that accent adaptation across speakers is more difficult than within one speaker, but is possible. Possibly for this reason, accent learning in the present study only occurred for items that were included in the training phase (old tokens) and did not generalize to new words. In line with abstractionist accounts of the mental lexicon (e.g., [16, 17]), we assume that the amount of training was not enough for full abstraction. Intensifying the training would probably have evoked training effects also for new tokens. Note however, that in prior research [3] single Dutch words with both a global and a specific Hebrew accent induced priming effects after only 3.5 minutes of phoneme monitoring training. However, unlike the present study and [8], [3] presented the same speaker during training and test.

Interestingly, learning effects did not differ between the production and the listening groups; L1 English participants learned an accent in their L1 equally well with both production and listening training. This finding is in line with the results in [10]. The eye-tracking study with single accented German words revealed similar proportions of looks to the target by L1 German participants after production and after listening training. Why then was there a production advantage for L2 participants in [8]? This is probably because accent strength still plays a role in accent learning; accent strength modulates accent learning together with speaker-listener similarity.

The role of accent strength in accent adaptation with listening training is supported by [2]. Mild accents are more easily learned than stronger accents. Accent strength is further emphasized in the accent processing classification that was postulated in [1]. In this account, the accent's acoustic distance from native speech is the only decisive factor in accent adaptation. Foreign and native accents follow the same principles, but the strength of an accent determines the ease of accent adaptation. When learning through listening, stronger accents need more time or more intense training. In [8], the speakers were L2 speakers, which involves a stronger accent than in the present study, where L1 speakers were recorded. The L2 speakers featured both global accent markers and the specific, manipulated accent marker, whereas the L1 speakers in the present study only produced the specific accent marker, implying a smaller distance from canonical pronunciation.

4. Conclusion

Can differences in accent learning from L1 and from L2 speakers be explained by accent strength differences alone? The present results suggest that to probably not be true: Accent similarity between speaker and listener facilitates accent learning. Typically, an L1 user's accent is more similar to a second L1 user's accent than it is to that of an L2 user. This assigns an important role to the speaker's native language background. Still, the L1–L2 speaker comparison has shown that accent strength *per se* co-determines accent learning.

5. References

[1] C. M. Clarke and M. F. Garrett, "Rapid adaptation to foreign-accented English," *Journal of the Acoustical Society of America*, vol. 116, no. 6, pp. 3647–3658, 2004.

[2] M. J. Witteman, A. Weber, and J. M. McQueen, "Foreign accent strength and listener familiarity with an accent codetermine speed of perceptual adaptation," *Attention, Perception, & Psychophysics*, vol. 75, no. 3, pp. 537–556, 2013.

[3] M. J. Witteman, N. P. Bardhan, A. Weber, and J. M. McQueen, "Automaticity and stability of adaptation to a foreign-accented speaker," *Language and Speech*, vol. 58, no. 2, pp. 168–189, 2015.

[4] B. G. Evans and P. Iverson, "Plasticity in vowel perception and production: a study of accent change in young adults," *Journal of the Acoustical Society of America*, vol. 121, no. 6, pp. 3814–3826, 2007.

[5] A. M. Trude and S. Brown-Schmidt, "Talker-specific perceptual adaptation during online speech perception," *Language and Cognitive Processes*, vol. 27, no. 7-8, pp. 979–1001, 2012.

[6] F. Eisner, A. Melinger, and A. Weber, "Constraints on the transfer of perceptual learning in accented speech," *Frontiers in Psychology*, vol. 4, 2013.

[7] M. J. Witteman, A. Weber, and J. M. McQueen, "Tolerance for inconsistency in foreign-accented speech," *Psychonomic Bulletin & Review*, vol. 21, no. 2, pp. 512–519, 2014.

[8] A.-K. Grohe and A. Weber, "Learning to comprehend foreign-accented speech by means of production and listening training," *Language Learning*, in press.

[9] A. Hanulíková and A. Weber, "Sink positive: linguistic experience with th substitutions influences nonnative word recognition," *Attention, Perception, & Psychophysics*, vol. 74, no. 3, pp. 613–629, 2012.

[10] A.-K. Grohe and A. Weber, "The Penefit of salience: salient accented, but not unaccented words reveal accent adaptation effects," *Frontiers in Psychology*, 2016.

[11] R. H. Baayen, R. Piepenbrock, and L. Gulikers, The CELEX lexical database (Release 2). Philadelphia: University of Pennsylvania, 1995.

[12] R. H. Baayen, D. J. Davidson, and D. M. Bates, "Mixed-effects modeling with crossed random effects for subjects and items," *Journal of Memory and Language*, vol. 59, no. 4, pp. 390–412, 2008.

[13] A. R. Bradlow and T. Bent, "Perceptual adaptation to non-native speech," *Cognition*, vol. 106, no. 2, pp. 707–729, 2008.

[14] C. T. McLennan and J. González, "Examining talker effects in the perception of native- and foreign-accented speech," *Attention, Perception, & Psychophysics*, vol. 74, no. 5, pp. 824–830, 2012.

[15] S. Weil, "Foreign accented speech: encoding and generalization," *Journal of the Acoustical Society of America*, vol. 109, no. 5, p. 2473, 2001.

[16] J. L. McClelland and J. L. Elman, "The TRACE model of speech perception," *Cognitive Psychology*, vol. 18, no. 1, pp. 1–86, 1986.

[17] D. Norris, "Shortlist: A connectionist model of continuous speech recognition," *Cognition*, vol. 52, no. 3, pp. 189–234, 1994.

Patterns of gender variation in the speech of primary school-aged children in Australian English: the case of /p t k/

Casey Tait & Marija Tabain

La Trobe University, Melbourne, Australia

ceford@students.latrobe.edu.au, m.tabain@latrobe.edu.au

Abstract

This paper examines the pattern of gender-specific variation of voiceless plosives /p t k/ in the speech of primary school-aged children (ages 5-12) in Australian English. While results are preliminary, findings at this stage show gender-specific variation of the voiceless plosives in the speech of primary school children, even in the youngest age group examined.

Index Terms: Australian English; sociophonetics; gender variation; plosives.

1. Introduction

Several studies of Australian English have focused on variation in the production of the voiceless plosives /p t k/ [1]–[4]. However, none of that research addresses phonetic variation in the speech of Australian English-speaking children. Studies of gender-specific variation of these plosives in the speech of children in other varieties of English predominantly focus on children who are still largely in the language acquisition phase [5]–[7]. Research has shown that the speech of children prior to school-age seems to reflect the features in the speech of their primary caregiver e.g. [6], [8]. Once children reach school age, their speech becomes more like that of their peers, and becomes more so as they move toward adolescence [9]. Adolescence marks the age at which speakers are considered to have reached a level of linguistic mastery, as well as the onset of puberty and the development of gender-specific differences in the vocal tract anatomy. The primary school years are therefore the transition period where children move from home-based care into wider society and into adolescence. Examining this age period could give us important insight into the development of gender-specific phonetic variation prior to adolescence.

This paper presents preliminary patterns of gender-specific variation of voiceless plosives /p t k/ in Australian English by primary school-aged children. In particular, instances of glottalisation, tapping, frication, and pre-aspiration were analysed, as studies of /p t k/ in adult speech in Australian and other varieties of English have found these variants to be particularly marked for gender.

Of the voiceless plosives set, variation of /t/ has received the most attention. Loakes and McDougall's [1] work on plosive frication notes that fricated /t/ appears very infrequently in the speech of their male participants, suggesting that their avoidance of this variant is due to its association with female speakers. This notion has been reinforced by studies of /t/ that include solely female participants by its greater frequency of use [2], [10]. The frication or affrication of /t/ is also noted to be a particularly female variant as it is in other varieties of English such as Tyneside English [11] and New Zealand English [12].

Pre-aspiration of /t/ has not been fully investigated in Australian English, however in other varieties of English pre-aspiration of /t/ is predominantly a female feature, especially for younger females [13], [14].

The glottalisation of /t/ is noted to be used by Australian English speakers [2], however it has been investigated to a lesser extent than other variants in terms of gender variation. One small-scale study [4] found a negligible gender correlation in its use: males produced more glottal /t/ in read speech, females produced more in word list data, and conversational data showed a relatively even production of glottalisation between the two genders. Other varieties of English that have investigated /t/ glottalisation have also found its use to vary between the sexes depending on context within an utterance or class status [15]–[17].

The tapped variant of /t/ is noted in Horvath's 1985 work on Sydney English to be 'clearly male', and associates a 'heavily aspirated' variant with female speakers [3, p. 104].

Studies in the variation of /p/ and /k/ are somewhat fewer in Australian English. Loakes and McDougall [1] found that frication of /p/ and /k/ by their male participants occurs at a much higher rate than /t/, with fricated /k/ appearing most often. This finding was reinforced by a smaller study [4] of /p/ and /k/ variation, where males were found to fricate at higher rate than females. These phonemes have received more attention in other varieties of English, where the frication of /p/ and /k/ is also reported as a predominantly male feature [14], [16], [17]. Similarly, the pre-aspiration of /p/ and /k/, while not investigated in Australian English, has been associated with female speakers in other varieties of English [16].

There is currently no research that addresses plosive variation in the speech of Australian English-speaking children. There have been some findings in other varieties of English that focus on gender variation of plosives, but mostly for children before the age of four years [5]–[7]. These studies have found that these children are beginning to show sensitivity to gendered variation of plosives. Docherty and colleagues [7] found that pre-aspiration appeared predominantly in the speech of girls in their study, reflecting the pattern of use in the speech of the adults in the community. Milroy and colleagues [18] found that children around 5 years of age are beginning to produce gender-based variation in the realisation of glottalised stops in British English.

The current study therefore aims to build upon our knowledge of children's sociophonetic development in the years between the language acquisition phase and adolescence, particularly in Australian English.

2. Method

2.1. Participants

The data presented here are preliminary, and represent a subset of a larger corpus collected by Casey Tait as part of a doctoral thesis on the acquisition of sociophonetic cues to gender in the speech of Australian English-speaking primary school-aged children.

The use of the different variants of /p t k/ are examined in the speech of 18 Australian English-speaking children from primary school in Yarrawonga, Victoria. Students from three year levels were recorded: Prep (ages 5-6), Year Three (ages 8-9), and Year Six (ages 11-12). This subset contains 3 boys and 3 girls from each year level, making a total of 9 girls and 9 boys. These particular year levels were selected in order to facilitate a developmental overview of the seven-year primary school period.

2.2. Recording

Speakers were recorded in same-sex and year level-matched dyads and were recorded on school grounds during school hours. The Year Six recordings took place at a separate campus to the Year Prep and Year Three recordings. All recordings were made using a Marantz Professional PMD661 solid state recorder and Shure SM-94 microphones at a sampling rate of 44.1kHz. Each speaker had a separate microphone placed in front of them on a boom stand. Microphones were placed at around 20-30cm from the speaker's mouth (or as close to this as was possible). The recording sessions lasted between 30-60 minutes and involved a mixture of spontaneous conversation and interactive games and activities.

2.3. Labelling

Each speaker's recording was broken down into shorter, manageable sections, transcribed, and then segmented and force-aligned using the WebMAUS-multiple automated alignment service [19]. Segment boundaries and consonant labels were manually adjusted using the *EMU speech database system* [20].

Fricated tokens were identified acoustically by an area of high frequency energy and lack of stop closure and release. An affricated variant of /t/ was also identified and considered here under the /t/ 'frication' category. These were auditorily distinct from a canonical released /t/ and were identified acoustically with a stop closure and a long period of aspiration that began with a short period of higher energy.

Pre-aspirated variants were identified by a period of fricative energy preceding the stop closure.

The glottalised variant of /t/ contained two realisations: the first (glottal /t/) was identified at areas where there were no formant transitions and the presence of creaky phonation on either side of the stop closure, while the second (laryngealised /t/) was identified by a lack of stop closure or release and the presence of fully laryngealised voicing throughout the segment.

Taps were identified acoustically by a short closure phase and a short period of voicing.

3. Results

3.1. Occurrence of variables

Table 1 shows the overall occurrence of the /p t k/ variables for boys and girls in each of the three year levels. Tokens were counted in four utterance environments: word-medial intervocalic, word-final intervocalic, word-final pre-consonantal, and word-final pre-pausal. On average, each girl produced around 8 minutes of speech from the four activities, while each boy produced around 14 minutes of speech. This accounts for the difference in the overall number of tokens between the boys and girls. Even though there is a large difference in the overall occurrence of each variable between the two gender groups, there are some patterns of gender-specific variation of these three voiceless plosives that are emerging.

	/t/ tokens		/p/ tokens		/k/ tokens	
Year Level	**Boys**	**Girls**	**Boys**	**Girls**	**Boys**	**Girls**
Prep	314	179	74	67	124	52
Three	514	281	115	32	258	82
Six	716	307	132	50	299	111
Total	**1544**	**767**	**321**	**149**	**681**	**245**

Table 1: *Total token counts for /p t k/ for 18 speakers, broken down by sex and year level. Each speaker group contains 3 speakers.*

3.2. /t/ variation

3.2.1. Overall patterns of gender-specific /t/ variation

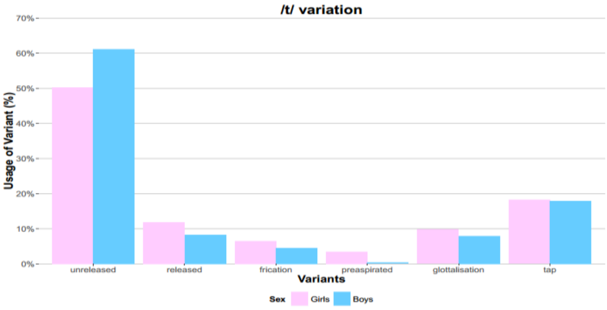


Figure 1: *Rate of use for each /t/ variant by boys and girls overall (Boys n=9; Girls n=9).*

Figure 1 shows the rate of use for each variant of /t/ for boys and girls overall. Both the fricated (including both the fully fricated and affricated realisations) and the pre-aspirated variants of /t/ appear in the girls' speech at a higher rate than boys, which is consistent with findings of these variants in adult speech. Glottalisation of /t/ (including both the glottal and laryngealised realisations) occurred at a higher rate in the speech of girls than boys overall. The tapped variant appeared at around the same rate for both boys and girls.

3.2.2. Developmental patterns of /t/ variation

Figure 2 shows the rate of use for each /t/ variant by the six speaker groups examined. For the girls, there is a decrease in use of both the fricated and pre-aspirated variants of /t/ with an increase in age, but an increase in the use of both the glottalised and tapped variants with an increase in age. For the boys, there is a decrease in the use of the frication category of /t/ with an increase in age. An increase in the use of taps with

Table of Contents for this manuscript

an increase in age is apparent, with both the Year Three and the Year Six groups using tapped /t/ at around the same rate. For the glottalisation of /t/, there is an increase in use from the Prep to the Year Three group, and then a decrease by the oldest age group. Pre-aspiration of /t/ appears very rarely in each of the boy speaker groups.

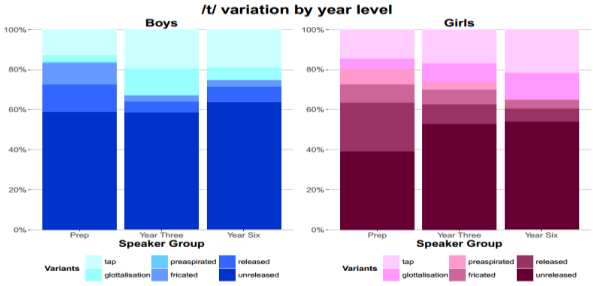


Figure 2: *The rate of use for each /t/ variant by each speaker group. Each group includes 3 speakers.*

3.3. /p/ variation

3.3.1. *Overall patterns of gender-specific /p/ variation*

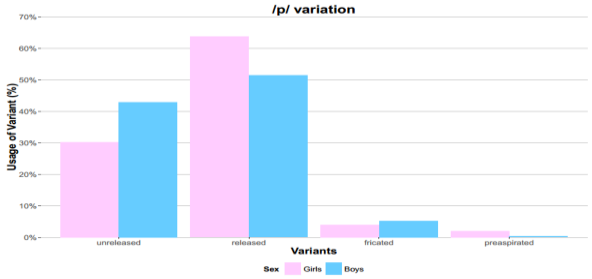


Figure 3: *Rate of use for each /p/ variant by boys and girls overall (Boys n=9; Girls n=9).*

Figure 3 shows the rate of use of each variant of /p/ by boys and girls overall. While there were considerably fewer tokens of /p/ produced overall compared to /t/ and /k/, there are still some gender-specific patterns emerging. The majority of /p/ tokens for both boys and girls are canonical released tokens of /p/, followed by an unreleased variant. The fricated variant of /p/ appears at a higher rate in the speech of boys than it does for girls. While the pre-aspirated variant of /p/ was relatively rare for both gender groups, it appeared more often in girls' speech than in boys' speech. Both of these findings are consistent with the use of these variants by adult speakers of both Australian English and other varieties of English.

3.3.2. *Developmental patterns of /p/ variation*

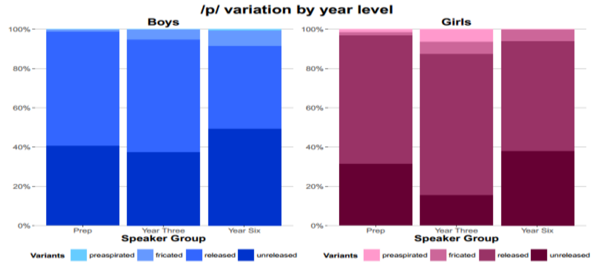


Figure 4: *The rate of use for each /p/ variant by each speaker group. Each group includes 3 speakers.*

Figure 4 shows the rate of use for each /p/ variant by the six speaker groups examined. For the fricated variant of /p/ there is an increase in use with an increase in age for boys. There is a similar pattern in its use for girls, with both the Year Three and Year Six age group produced a fricated /p/ at around the same rate. The pre-aspirated variant of /p/ increases in occurrence with an increase in age for girls from Prep to Year Three, but it does not appear at all in the speech of the oldest girls. For boys, the pre-aspirated variant only appears in the speech of the boys in the Year Six group, however it only makes up less than 1% of all of their /p/ tokens overall.

3.4. /k/ variation

3.4.1. *Overall patterns of gender-specific /k/ variation*

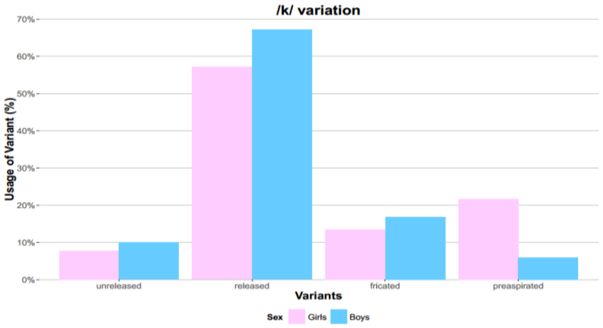


Figure 5: *Rate of use for each /k/ variant by boys and girls overall (Boys n=9; Girls n=9).*

Figure 5 shows the rate of use of each variant of /k/ by boys and girls overall. The majority of /k/ variants were released canonical tokens. Again, while there were considerably fewer tokens of /k/ produced by both gender groups in comparison to /t/, some gender specific patterns are apparent for this variable in the children's speech. The frication and pre-aspirated variants of /k/ patterned in a very similar way as the fricated and pre-aspirated variants of /p/: /k/ frication appeared at a higher rate in the speech of boys than in the speech of girls, while the pre-aspirated variant of /k/ is more preferred by girls overall than boys. Again, these findings are consistent with previous findings of the variation of /k/ in adult speech.

3.4.2. *Developmental patterns of /k/ variation*

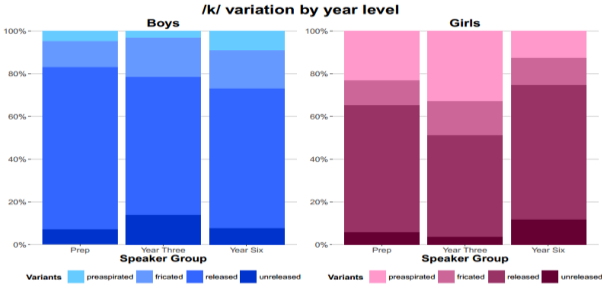


Figure 6: *The rate of use for each /k/ variant by each speaker group. Each group includes 3 speakers.*

Figure 6 shows the rate of use of each variant of /k/ by the six speaker groups examined. The frication of /k/ appears at a higher rate in the boys' speech in all three year levels, but for both gender groups there seems to be an overall increase in use with an increase in age. For girls, the pre-aspirated variant of /k/ increases in occurrence from the Prep to the Year Three group, but a decrease in use occurs in the movement from the Year Three to the Year Six group. For boys, there is an overall pattern of increase in pre-aspirated /k/ with an increase in age, although it appears at a much lower rate than for girls at all year levels.

4. Discussion

While the results are preliminary, patterns of variation of the voiceless plosives /p t k/ that correlate with gender are evident in the speech of primary school children in Australian English. Overall, girls are showing evidence of using variants that are largely associated with female speakers at a higher rate than boys. In particular, the pre-aspiration of all three of the voiceless plosives is highly preferred by girls, as well as the frication of /t/. Similarly, boys produced higher rates of male-correlated variants than girls, particularly the frication of /p/ and /k/. It is important to note that even speakers in the youngest age groups examined (ages 5-6 years) are producing variants that correlate with their gender.

Tollfree's evaluation of /t/ variation noted speakers' awareness of a formality hierarchy of /t/ variants [2, p. 60]. In particular, her speakers associated a fricated realisation with more formal or 'correct' speech, while they associated the tapped variant with more casual and laid back speech. For both boys and girls, there decrease in the use of fricated /t/ realisations, and an increase in the use of the tapped variant. The results reported here seem to suggest that children are becoming more aware of this formality hierarchy as they move toward adolescence.

It is also possible that the use of more 'casual' variants is associated with speech used in the community, or in regional areas of Victoria (or Australia) in general, and that children are becoming more aware of this association as they grow in age. This idea may be further extended when examining the variants used by the girls closest to adolescence. Overall, their use of variants patterns quite closely with boys in the same age group. For all female-correlated variants (i.e. pre-aspiration, /t/ frication), there is a decrease in their rate of use in comparison to girls in the younger age groups, while they employ the more male-correlated variants such as /p/ and /k/ frication at a similar rate as the Year Six boys.

The relationship between the variation of segmental features and social membership is well established in sociolinguistic theory (e.g. [21]). It may be possible that the use of more male-correlated, or possibly 'non-standard', phonetic variants and sociolinguistic variables could be a feature of the speech of the community. The analysis of further data and additional variables is planned for future work which may provide further insight into the patterns of plosive variation that have been revealed here. Furthermore, statistical significance testing is planned which may also provide further clarity in regards to the strength of each of the variables analysed. Mixed effects models may also reveal individual speaker differences that could be at play.

5. Acknowledgements

Thanks go to the students at Yarrawonga College P-12 who participated in this study, their parents and guardians, and their classroom teachers. Particular thanks go to Meagen Rae and Kim Stewart for their assistance and patience in the organisation of the data collection. Thanks also go to those who have commented on this work at its various stages. Your feedback and support is much appreciated.

6. References

[1] D. Loakes and K. McDougall, "Individual Variation in the Frication of Voiceless Plosives in Australian English: A Study of Twins' Speech," *Aust. J. Linguist.*, vol. 30, no. 2, pp. 155–181, May 2010.

[2] L. Tollfree, "Variation and change in Australian consonants: reduction of /t/," in *Varieties of English Around the World: English in Australia*, D. Blair and P. Collins, Eds. Amsterdam: John Benjamins Publishing Co., 2001, pp. 45–67.

[3] B. Horvath, *Variation in Australian English. The sociolects of Sydney*. Cambridge: Cambridge University Press, 1985.

[4] V. W. Y. Su, "The Gender Variable in Australian English Stop Consonant Production," Unpublished Honours Thesis, University of Melbourne, 2007.

[5] P. Foulkes, G. Docherty, G. Khattab, and M. Yaeger-Dror, "Sound Judgements: Perception of Indexical Features in Children's Speech," in *A Reader in Sociophonetics*, D. R. Preston and N. Niedzielski, Eds. New York: Walter de Gruyer, 2010, pp. 327–356.

[6] P. Foulkes, G. Docherty, and D. Watt, "Phonological Variation in Child-Directed Speech," *Linguist. Soc. Am.*, vol. 81, no. 1, pp. 177–206, 2005.

[7] G. Docherty, P. Foulkes, B. Dodd, and L. Milroy, "The emergence of structured variation in the speech of Tyneside infants," *Final Rep. ESRC Award #R000237417*, pp. 12–31, 2002.

[8] H. Buchan and C. Jones, "Phonological reduction in maternal speech in northern Australian English: change over time.," *J. Child Lang.*, vol. 41, no. 4, pp. 725–55, Jul. 2014.

[9] P. Kerswill, "Children, adolescents, and language change," *Lang. Var. Change*, vol. 8, no. 2, pp. 177–202, 1996.

[10] M. J. Jones and K. McDougall, "The acoustic character of fricated /t/ in Australian English: A comparison with /s/ and /ʃ/," *J. Int. Phon. Assoc.*, vol. 39, no. 3, pp. 265–289, 2009.

[11] G. Docherty and P. Foulkes, "A corpus-based account of variation in the realisation of 'released' /t/ in English," in *Proceedings of the 6th Australian International Conference on Speech Science and Technology*, 1996, pp. 157–162.

[12] G. Docherty, J. Hay, and A. Walker, "Sociophonetic patterning of phrase-final /t/ in New Zealand English," in *Proceedings of the 11th Australian International Conference on Speech Science & Technology*, 2006, pp. 378–383.

[13] P. Foulkes and G. Docherty, "The social life of phonetics and phonology," *J. Phon.*, vol. 34, no. 4, pp. 409–438, Oct. 2006.

[14] P. Honeybone, "Lenition inhibition in Liverpool English," *English Lang. Linguist.*, vol. 5, no. 2, pp. 213–250, Sep. 2001.

[15] G. J. Docherty and P. Foulkes, "Sociophonetic Variation in 'glottals' in Newcastle English," *14th Int. Congr. Phonetic Sci.*, pp. 1037–1040, 1999.

[16] G. J. Docherty and P. Foulkes, "Derby and Newcastle: instrumental phonetics and variationist studies," in *Urban Voices: Accent Studies in the British Isles*, P. Foulkes and G. Docherty, Eds. London: Arnold, 1999, pp. 47–71.

[17] G. J. Docherty, P. Foulkes, J. Milroy, L. Milroy, and D. Walshaw, "Descriptive adequacy in phonology: a variationist perspective," *J. Linguist.*, vol. 33, no. 2, pp. 275–310, Sep. 1997.

[18] J. Milroy, L. Milroy, S. Hartley, and D. Walshaw, "Glottal stops and Tyneside glottalization: Competing patterns of variation and change in British English," *Language Variation and Change*, vol. 6, no. 3. p. 327, 1994.

[19] 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.

[20] J. Harrington, *Phonetic Analysis of Speech Corpora*. Malden, MA: Blackwell, 2010.

[21] W. Labov, "The social motivation of a sound change," *Word*, vol. 19, pp. 273–309, 1963.

Change in Māori focus/topic *ko*: the impact of language contact on prosody

Sasha Calhoun, Naoko Yui, and Karena Kelly

Victoria University of Wellington

sasha.calhoun@vuw.ac.nz, karena.kelly@vuw.ac.nz

Abstract

This paper looks at the impact of contact with English on the prosody of te reo Māori. We investigate the functions and prosody of sentences with pre-predicate focus/topic marker *ko* in three groups of present-day male speakers from the MAONZE corpus (King et al. 2010): older L1, and younger L1 and L2 speakers. Older speakers show the expected pattern of main stress in the *ko*-phrase for focus *ko*, and in the main clause for topic *ko*. L1Y speakers show a similar pattern, but with a weaker distinction. L2Y speakers primarily use topic *ko*, but apparently with "focus" prosody.

Index Terms: prosody, topic, focus, information structure, Māori, Austronesian, language change, language contact

1. Introduction

Language change is a natural phenomenon, occurring when languages meet each other [1]. It is believed that language contact can impact all levels of linguistic structure [1]; however, there have been relatively few studies looking at how language contact affects prosody (cf. [2]). This needs to be addressed, as prosody is an integral part of spoken language, signalling a wide range of meanings, including information structure [3].

Te reo Māori is the indigenous language of New Zealand, a Polynesian VSO language which came into contact with the English language in the early 19th century, leading to a period of dramatic language loss in the mid-20th century, and a subsequent language revitalisation movement from the 1980s to the present. National survey results in 2013 showed 50,000 people (11% of the adult Māori population) reported speaking Māori very well or well, around 14,500 of whom were aged 55+ years, and 18,000 were aged 25-44 (http://www.stats.govt.nz/browse_for_stats/people_and_communities/maori/TeKupenga_HOTP13/Tables.aspx).

There have been a few linguistic studies exploring the effect of contact with English on modern Māori. Kelly's [4] corpus-based investigation identified a number of aspects of syntactic change occurring between 1900 and 1990. The Māori and New Zealand English (MAONZE) project has investigated sound change in Māori, showing ongoing influence from English, as well as system-internal changes, using recordings of 60 male and female Māori speakers from three different generations with birth dates spanning over 100 years [5]. Thompson [6] looked at the perception of prosodic prominence among contemporary Māori and non-Māori speakers, using recordings from the MAONZE corpus.

The information structure of traditional Māori matches other Polynesian languages, and what has been proposed to be a rule for verb-initial languages, i.e. the focus is usually sentence initial, optionally preceded by a slot for establishing topics [7,8,9]. Foci present new information in relation to the current proposition and/or carry the discourse forward, whereas topics are the entities being discussed [10,11]. There are a number of constructions which place a focused constituent in initial position, including pre-predicate *ko*-phrases. The initial focus carries the main stress, which cannot generally be moved in the sentence to mark emphasis [7], ie. less intonational plasticity. By contrast, in English the focus tends to be final, and stress can be moved freely within a sentence to mark focus [3]. Anecdotal evidence suggests younger Māori speakers are more influenced by English means of marking information structure ([7]:218, [12]:664).

In this study, we look at the prosody of the focus marker *ko* (all examples from MAONZE):

(1)	Ko	tō	mana	kei	roto	kē	i	te	reo
	FOC	your	power	LOC	inside	INTENS	PREP	the	language

'Your power lies within the language.'

Here the speaker has just been talking about his language, so 'your mana' is new in the proposition, and *ko* marks focus.

However, *ko* can also be used to mark topics:

(2)	Ko	rātou	kei te	rongo	mai	i	te	haunga	ika
	TOP	they	TAM	sense	DIR	DO	the	odour	fish

'They are smelling the fishy odour.'

'They' were just mentioned in the previous discourse, so *ko* marks the topic. *Ko*-phrases can also appear after the predicate, but here only topic *ko* is attested ([12]:181). *Ko* has a number of other functions, eg. future locative preposition; these are not considered in this study.

It is claimed that in pre-predicate position, topic and focus *ko* are distinguishable prosodically, with the main stress falling within the *ko*-phrase for the focus usage, and outside the *ko*-phrase for the topic usage [10]. In Māori each phonological phrase is said to be marked with an H*L- pitch contour, with the H* associated with the lexically-stressed syllable of the head word [13]. Bauer ([12]:562) comments that the pitch peak in the first phonological phrase in a sentence can be raised to mark emphasis, however, the subsequent phrases are not dephrased or deaccented. We therefore expect the main stress to be marked by relative prominence, not accent placement or type.

While there have been anecdotal observations that the use of focus/topic *ko* is declining among younger Māori speakers, there has not been any research on this. Further, the prosodic distinction between topic v focus *ko* has not yet been examined using modern prosodic analysis, nor is it known if this distinction is maintained by younger speakers. This study sought to investigate these issues.

2. Method

2.1. Data

The data for this study was taken from the MAONZE corpus [5]. Two of the speaker groups have been analysed to date: present day male elders (kaumātua, Group K) and present day young males (Group Y) (recorded mid 2000s) (see summary Table 1). This intergenerational comparison provided a means to investigate language change. The interview data were used.

Group K speakers were raised speaking Māori in a rural Māori-speaking community, learning English after starting school ([5]:8). Group Y was divided into two sub-groups based on language experience: the L1Y sub-group were native speakers of Māori, raised in a Māori-speaking environment, and the L2Y sub-group were second-language speakers, who had learned Māori through schooling ([14]:318).

Table 1: *Summary of* ko *data*

	Speaker Group		
	K	**L1Y**	**L2Y**
Speakers	9	5	4
Year of birth	1925-38	1969-84	1969-84
Age at recording	64-79	21-35	21-35
Total *Ko* occurences	798	420	456
FOC/TOP *ko*	121	65	59
Pre-pred FOC/TOP *ko*	83	51	34

LaBB-CAT [15] was used to extract all sentences with *ko*-phrases (see Table 1). There were initially 10 K and 10 Y speakers, but one K and one L2Y speaker's recordings were excluded due to poor sound quality. The transcriptions for each recording were then exported as Praat Textgrids [16], with automatic word and phone alignments. Sentences with focus/topic *ko* were identified, glossed and translated into English. Surrounding discourse context was also translated to enable classification of each *ko* as marking topic or focus from the translation. Here we only consider pre-predicate *ko*.

2.2. Labelling topic and focus

The type and subtype of each focus/topic *ko* was then annotated following the guidelines in Skopeteas et al. [17]. All annotations were checked and agreed by all authors. Topic and focus were defined in section 1. Two subtypes of topic were identified: *aboutness topic* (TOPa), what the sentence is about (see (2)); and *contrastive topic* (TOPc):

(3)	Ko	tētahi	i	noho	mai	i	roto	o	Waipū
	TOP	one	TAM	stay	DIR	LOC	inside	of	Waipū

'One (of them) stayed in Waipū.'

The speaker's brothers were discussed previously, but he contrasts *tētahi* 'one of (them)' with the others.

Two subtypes of focus were also identified: *new-information focus* (FOCn), new and missing information which develops the discourse (see (1)); and *contrastive focus* (FOCc):

(4)	Ko	koe	anake	kei te	mōhio	ki	tērā
	FOC	you	alone	TAM	know	DO	that

'You alone know that.'

The Māori language (='tērā') is under discussion. '*Ko koe anake*' (only you) is contrasted with a semantically and/or syntactically parallel constituent in the discourse (i.e. other people) ([17]:172).

2.3. Prosodic analysis

To investigate a relative prominence distinction between focus and topic *ko* (cf. [10]), we compared the prosodic realization of the most prominent word in the *ko*-phrase, and that in the main clause. We manually segmented the word with the strongest accent in the *ko*-phrase and the main clause, including any preceding particle to capture the accentual rise, e.g. *ko tētahi* in the *ko*-phrase in (3). Accented words which were phrase final were labelled differently from non-phrase-final words, as the F0 contours of the former would be affected by the boundary tone. Mean acoustic measures and time-normalised F0 contours were automatically extracted for each labelled section using ProsodyPro tools [18].

3. Results

Table 2: *Frequency of* ko *subtypes by speaker group*

	Speaker Group			
Info	**K**	**L1Y**	**L2Y**	**Total**
FOCn	15.7%	5.9%	0.0%	16
FOCc	16.9%	17.6%	2.9%	24
TOPa	37.3%	33.3%	35.3%	60
TOPc	30.1%	43.1%	61.8%	68
Total	83	51	34	168

3.1. Frequency of focus/topic *ko*

The first analysis looked at the functions of *ko* for the different speaker groups (see Table 2). A logistic mixed effects model was built using the *lme4* package in *R* [19,20] with Info Type (TOP v FOC) as the dependent, Speaker Group as the fixed effect and Speaker as the random effect (N=168). This showed that all groups were significantly more likely to use TOP *ko* than FOC (Intercept p=0.002), however L2Y speakers were significantly more likely to use TOP than the others (p=0.008); there was no difference between K and L1Y.

We can therefore see that the older K speakers, and the younger L1Y speakers, use *ko* to mark both topics and foci; though focus *ko* is very rare for the L2Y speakers. Within the subtypes of FOC, it appears L1Y speakers are more likely to use FOCc function than FOCn, while K speakers use both functions equally (see Table 2). Within the subtypes of TOP, it appears L2Y speakers, and to a lesser extent L1Y speakers, use TOPc more than TOPa, while K speakers use both equally. However, these differences could not be verified using statistical tests because of the low counts.

3.2. Relative prominence of *ko*-phrase and main phrase

For traditional Māori (as spoken by the K speakers), it is claimed that for topic *ko* the main stress is in the main clause, whereas for focus *ko* it is in the *ko*-phrase (see section 1). Therefore, for focus *ko,* we should expect to see a drop in the acoustic correlates of stress between the *ko*-phrase and the main clause; but no drop for topic *ko*. We measured two key acoustic correlates of stress, mean F0 and intensity. In each case, the difference in the measure between the most prominent word in the *ko*-phrase and the main clause is taken (see section 2.2). As this is conversational data, with huge segmental variability, it was very difficult to measure durational differences.

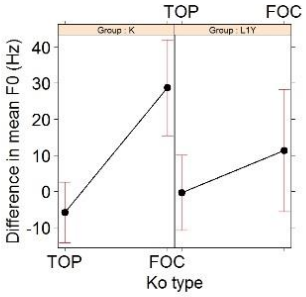


Figure 1: *Difference in mean F0 between the ko-phrase and main clause by ko Type and Speaker Group*

A linear mixed-effects model was built using the *lme4* and *lmerTest* packages [21] with the difference in mean F0 between the *ko*-phrase and the main clause as the dependent, Speaker Group (K v L1Y), *ko* Type (FOC v TOP) and their interaction as fixed effects, and Speaker as a random effect (N=120). The model showed a large drop in F0 when *ko* is a FOC (p<0.0001). The drop was much smaller for the L1Y speakers, although the difference between the groups only approached significance (p=0.067). An ANOVA comparing models with and without the interaction was marginally significant (p=0.057), but the data set was small. The model effects can be seen in Figure 1 (a larger difference indicates a greater fall in F0 between the *ko*-phrase and the main clause).

A linear mixed-effects model was built with the difference in mean intensity as the dependent, and the same fixed and random effects as above (N=132). This showed that there was a significant drop in intensity with FOC *ko*, compared to TOP *ko* (p<0.0001); however, there was much less of a drop for the L1Y speakers than the K speakers (p=0.021). An ANOVA showed that the model with the interaction between *ko* Type and Speaker Group was significantly better than one without (p=0.021). The model effects can be seen in Figure 2.

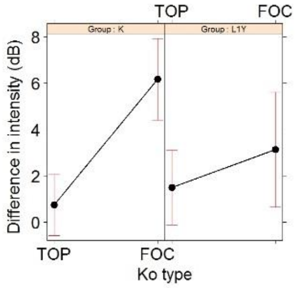


Figure 2: *Difference in mean intensity between the* ko-*phrase and main clause by* ko *Type and Speaker Group*

These results follow the expected pattern for the older K speakers: there is a large drop in mean F0 and intensity between the *ko*-phrase and the main clause for focus *ko*, consistent with the *ko*-phrase carrying the main stress. No drop is found for topic *ko*, consistent with the main stress occurring in the main clause. The L1Y speakers show the same pattern, but the F0 and intensity drops are much smaller for focus *ko*, and there is much more overlap in prominence for the two functions of *ko*. The L2Y speakers were not included as they had very few focus *ko* tokens. Interestingly, their differences for topic *ko* were roughly in between the values for topic and focus *ko* for the L1Y speakers (mean F0 difference=3.6Hz, sd=17Hz, outliers removed; mean intensity difference=1.8dB, sd=2.9dB).

3.3. F0 contours of *ko-phrase* and main clause

The last part of our analysis is exploratory. We wished to consider the F0 contour of the *ko*-phrase and the main clause as a whole, to see how the relative prominence differences found above manifest themselves for different groups, and if any of the sub-types of topic or focus are being distinguished.

Figure 3 (next page) shows time-normalised F0 contours for the most prominent word in the *ko*-phrase and the main clause, where these were phrase-final, produced by the ProsodyPro tools (see section 2.3. Non-phrase-final words are not shown for space reasons, as the phrase-final word would have to be shown separately). For the topic *ko*-phrases, sub-types are shown (blue dashed lines for TOPa, solid red for TOPc). There were no clear differences between FOCn and FOCc for the K group, and very few tokens for the L1Y group, so these are grouped.

For the K speakers, although there is variation, we can see that the contour for topic *ko* is generally either flat or rising (Fig. 3ai), while focus *ko* starts high and falls (Fig. 3bi). The main clause generally has a clear rise-fall when it is focal (Fig. 3aii), but a flat contour at a lower F0 when the *ko*-phrase is focal (Fig. 3bii). The L1Y speakers show a similar pattern for the *ko*-phrases (Figs. 3ci&di). However, apart from a few exceptions, the main clause has a flat, low pattern whether it carries the focus or not (Fig. 3cii&dii), although the drop is greater when the *ko*-phrase is focal. For the L2Y speakers, the F0 contour for the *ko*-phrase is falling, i.e. similar to the focus pattern for the other groups (Fig. 3ei); there is also an F0 drop in the main clause, although there is also often a rise-fall accent (Fig. 3eii). Interestingly, the L1Y speakers seem to distinguish TOPa and TOPc sub-types, i.e. TOPa contours are usually low and flat, while TOPc are higher and rising (Fig. 3ci). K speakers have a mix of flat and rising contours for topic *ko* as well, but these do not consistently map to the aboutness/contrastive distinction. The function of these for the K group remains to be explored.

4. Discussion

The results of this study are consistent with the descriptions of focus/topic *ko* in the literature for the older male speakers (K). When *ko* marks focus, there is a clear drop in prominence between the *ko*-phrase and the main clause, consistent with the main stress being in the *ko*-phrase; whereas for topic *ko* the strongest accent in the main clause is equally or more acoustically prominent than that in the *ko*-phrase, consistent with this carrying the main stress. For this generation of speakers, the organization of information fits the pattern expected for Polynesian and other verb-initial languages.

There were clear differences among the young male speakers depending on whether they were first (L1Y) or second (L2Y) language speakers of Māori. L1Y speakers broadly followed the patterns for K speakers, however, there were signs of change consistent with greater contact with English. Focus *ko* was usually contrastive for these speakers, and contrastive topics were more common than aboutness, so *ko* may be developing into a marker of contrast. The focus position in English is usually final, however, foci can be initial if they are contrastive [11]. There is a relative prominence distinction between topic and focus *ko*, but is much weaker. Interestingly, there are signs that L1Y speakers may be developing distinct contours to mark subtypes of topics: flat for aboutness, rising for contrastive. This type distinction is salient in English. The K speakers appear to have these two topic contours as well, but we have not yet established the functional distinction between them. It may be that their functions have been remapped for the L1Y generation, similar to what Queen observed for Turkish-German speakers [2]. The L2Y speakers primarily use *ko* to mark topics. This is consistent with the influence of their L1 English focus-final pattern. Notably, however, the prosodic pattern for their topic *ko* more closely resembles focus *ko* for the other speaker groups: a falling contour on the *ko*-phrase and lower pitch in the main clause. It may be that these speakers have acquired the more phonetically salient focus *ko* prosodic contour, but applied it to the more familiar topic function.

This investigation of focus/topic *ko* has given us a window into information structure in Māori and its prosodic realization, and how this is being influenced by contact with English over time. The area is fascinating, as traditional Māori and English have opposite information ordering strategies and differing intonational plasticity. The research is just beginning: to date we have only considered one focus/topic marking construction with a limited data set. The prosodic analysis is also exploratory. The kinds of prosodic contour distinctions discussed in section 3.3 may be differences in gradient prominence or phonological tonal type. However, the evidence is that the impact of English contact is complex, with the same syntactic and prosodic resources used for different functions between generations of Māori speakers.

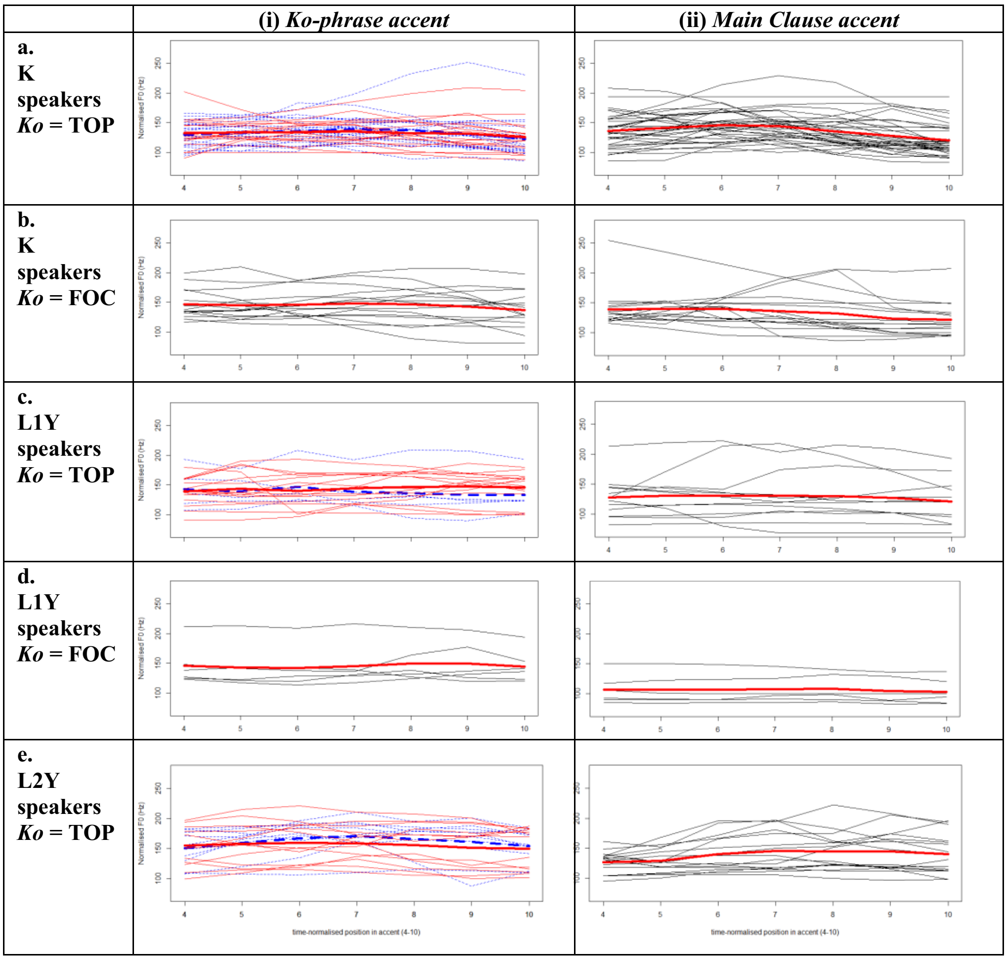


Figure 3: *Time-normalised F0 contours for phrase-final accented words in the* ko*-phrase and main clause by Speaker Group and* ko *Type. F0 level was sampled at 10 equally spaced points, points 4-10 are shown. The thick line (red or blue) shows the mean over all contours. For the TOP contours, dashed blue lines show TOPa, and solid red TOPc. Pitch tracking errors were removed.*

5. Acknowledgements

Mei kore ake te tīma MAONZE, i ora ai te kaupapa nei, me te karahipi a VUW hei tautoko i a Naoko.

6. References

[1] Sankoff, G. 2001, Linguistic Outcomes of Language Contact. In J.K. Chambers, P. Trudgill & N. Schilling-Estes (eds), *Handbook of Language Variation and Change*. UK: Blackwell, 638-668.

[2] Queen, R. 2012. Turkish-German bilinguals and their intonation: Triangulating evidence about contact-induced language change. *Language*, *88*(4), 791–816.

[3] Ladd, D. 2008. Intonational phonology (2nd ed), UK:CUP

[4] Kelly, K. 2015, *Aspects of change in the syntax of Māori – a corpus based study*, Ph.D. thesis, Victoria University of Wellington.

[5] King, J., Maclagan, M., Harlow, R., Keegan, P. and Watson, C. 2010. The MAONZE corpus: Establishing a corpus of Maori speech, *New Zealand Studies in Applied Linguistics* 16(2):1-16.

[6] Thompson, L. 2015. *Eliciting and analysing perceptions of prosodic prominence: a Māori case study*, PhD thesis, U. of Auckland.

[7] Bauer, W. 1993. *Maori*, Routledge, London.

[8] Calhoun, S. 2015. The interaction of prosody and syntax in Samoan focus marking. *Lingua* 165: 205-229.

[9] Herring, S. 1990. Information structure as a consequence of word order type, in *Proc. of BLS* (pp 163-174).

[10] Bauer, W. 1991, Maori *ko* again, *Te Reo* 34:3-14.

[11] Gundel, J. & Fretheim, T. 2005. Topic and Focus. In R. H. Laurence & G. Ward (eds), *The Handbook of Pragmatics*, UK: Blackwell, 175-196.

[12] Bauer, W. 1997. *The Reed reference grammar of Māori*, Reed, Auckland.

[13] De Lacy, P. 2003. Constraint universality and prosodic phrasing in Māori. In A. Carpenter, A. Coetzee & P. de Lacy (eds), *Papers in Optimality Theory II*, GLSA, 59-79.

[14] Watson, C., MacLagan, M., King, J., & Harlow, R. 2006. Are there L1 and L2 effects in the speech of young speakers of Māori?. In *Proc. of SST* (pp 317-322). Auckland.

[15] Fromont, R. & Hay, J. 2012. LaBB-CAT: an Annotation Store. http://www.aclweb.org/anthology/U12-1015

[16] Boersma, P. & D. Weenink 2016. Praat: doing phonetics by computer [Computer program], http://www.praat.org/.

[17] Skopeteas, S., Fiedler, I., Hellmuth, S., Schwarz, A., Stoel, R., Fanselow, G., Féry, C., Krifka, M., 2007. Questionnaire on information structure: reference manual. In Ishihara, S., Schmitz, M. (eds), *Working Papers of the SFB 632*.

[18] Xu, Y. 2013. ProsodyPro – A Tool for Large-scale Systematic Prosody Analysis", in *Proc. of TRASP 2013*, Aix-en-Provence.

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

[20] R Core Team 2016. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/

[21] Kuznetsova, A., Brockhoff, P., and Christensen, R., 2016. lmerTest: Tests in Linear Mixed Effects Models. R package version 2.0-30. https://CRAN.R-project.org/package=lmerTest

Disambiguation of Australian English vowels

Tunde Szalay, Titia Benders, Felicity Cox, Michael Proctor

Department of Linguistics, Macquarie University, Sydney, Australia

{tunde.szalay, titia.benders, felicity.cox, michael.proctor}@mq.edu.au

Abstract

Discriminability of Australian English vowel pairs was examined using a lexical decision task. 15 native listeners categorized 10,800 Australian English words with /hVd/ structure in a binary forced choice task. 16 target words differing only in the nuclear vowel were presented as auditory stimuli, paired with each of the remaining 15 words as competitors. Accuracy and reaction time were measured. Results show that target vowel and the phonetic similarity of target vowel and competitor vowel affect lexical decision.

Index Terms: perception, lexical decision, Australian English, vowel discriminability

1. Introduction

A large body of research has examined word recognition, and the factors that are involved. Lexical frequency [1], familiarity [2], morpho-syntactic context [3, 4], and neighbourhood effects – both semantic [5] and phonological [6, 7, 8] – have been shown to influence the speed and accuracy with which an auditory lexical stimulus is classified. Many other factors are also involved in word recognition in non-auditory domains [9, 10].

Lexical access has been shown to be sensitive to within-category gradient variation in phonetic factors including VOT [11, etc.]. Listeners make different decisions according to the amount of information that is available [12, 13, 14], and these studies showed how and when phonemes are identified and disambiguated from each other.

While much is known about mechanisms of lexical disambiguation in American English, fewer studies have examined the details of these processes in Australian English (AusE). Most of this work has focused on perception at the level of the word and the segment [15]. AusE presents special challenges to the listener because it uses a large vowel inventory containing 18 stressed vowels (plus schwa) and incorporating phonemic vowel length contrast for certain spectrally similar pairs. For example, pairs of vowels such as /ɐ-ɐː, e-eː/, and /ɪ-iː/ have similar spectral quality but contrast in length. AusE also has monophthong-diphthong pairs such as /æ-æɪ, ɑe-ɔ/ and /æ-æɔ/ in which the monophthong is related to the first element of the diphthong, and /ʉː-əʉ, ɔ-æɔ/, in which the monophthong is related to the second element of the diphthong [16].

Such a vowel space makes intrinsic vowel similarities and vowel discriminability important.To better understand how listeners of AusE discriminate vowels in a rich and temporally-differentiated vowel space, we used a perception experiment to examine confusability. The aim of this experiment was to discover which vowels and vowel pairs are intrinsically hard to disambiguate.

2. Method

Vowel disambiguation was examined using a binary forced-choice lexical decision task [18]. Participants were presented with an auditory stimulus, and asked to identify the word by selecting one of two candidates presented orthographically on a computer screen. Participants were instructed to select the word they heard as quickly as they could.

2.1. Participants

15 native speakers of Australian English (14 female; ages 19 to 47; mean 22.5 years) took part in the experiment. All were undergraduate students of linguistics at Macquarie University who received class credit for participation. Fourteen participants were born in Australia and one immigrated before the age of two. 46% were monolinguals; other languages spoken by participants were Danish, Italian, Japanese, Korean Spanish, Teochew, and Vietnamese. One participant was left-handed.

2.2. Stimuli

The stimulus set consisted of recordings of single word utterances of the form /hVd/. 9 words and 7 non-words contrasted all the stressed vowels of Australian English other than /ɪə/ and /eː/. The informant was a 21 y.o. monolingual female university student, born in Sydney to Australian-born parents. All stimuli were recorded at 44.1 KHz, amplitude-normalized, truncated to common temporal landmarks, and digitized as 16 bit WAV files.

Stimuli were also presented orthographically to elicit participants' responses to audition. All words were presented in the form <hVd> to maintain orthographic regularity across items [9] and to avoid multi-morphemic representations [3, 4]: /hɜːd/, /hæɔd/, /hʉːd/ were presented as *herd*, *howd*, *hude* (avoiding *heard*, *how'd*, *who'd*). Non-words were represented with transparent regular spellings and piloted with native speakers of AusE.

2.3. Procedure

Participants were seated in a semi-enclosed booth wearing Sennheiser HD 380 pro headphones, facing a computer monitor. Visual stimuli were presented using E-Prime [19] software on an Asus laptop (60 Hz screen refresh). Participants responded by pressing one of two labelled buttons on the button box response tool to log response accuracy and RT.

The procedure consisted of three phases: a familiarisation phase, a practice test, and an experimental phase. During the familiarisation phase, a single word was presented orthographically for 4000 ms and the same word was presented auditorily

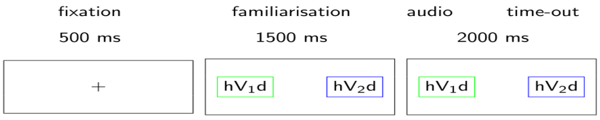


Figure 1: *Schematic structure of a trial in the test phase. Listeners selected the target word by pressing the corresponding button on the button box.*

at the midpoint of the visual presentation interval (at 2000 ms).

During the practice test, a non-word target and a non-word competitor were presented on the screen for 3000 ms. Audio presentation of the target word commenced at 1000 ms, allowing participants 2000 ms from the beginning of the auditory stimulus to make their choice. If they selected the target word correctly, they received the feedback "correct". If they gave an incorrect answer or did not provide an answer within 2000 ms, they received the feedback "incorrect" or "too slow" and the trial was repeated. The aim of the practice test was to strengthen the link between the spelling and pronunciation of non-words before the experimental phase.

During the experimental phase, participants first saw a fixation cross for 500 ms. The cross was followed by two words presented orthographically for 1500 ms, allowing participants time to read the words. 1500 ms after the two candidate words appeared on the screen, audio presentation of the target word began. Participants were allowed 2000 ms from the beginning of the auditory stimulus to identify the word by pressing a button on the same side as the corresponding orthographic representation. Participants were instructed to respond as quickly as possible. Figure 1 shows the schematic representation of trials in the experimental phase. If participants responded, the experiment moved on to the next trial without a feedback. If they did not respond within 2000 ms, they received the feedback "too slow" and the experiment only moved on to the next trial when the participant indicated that he/she was ready. The experimental phase was repeated three times, giving $3 \times 240 = 720$ trials.

2.4. Data analysis

We collected accuracy and reaction time (RT) data from the experimental phase. Accuracy and RT data were examined to determine whether any outliers should be removed. Participants' mean accuracy was 98.78% (range: 93%–100%), therefore no participant was excluded on the basis of accuracy. The mean RT of each participant fell within 2 standard deviations of the grand mean RT across all participants (644.8 ms), therefore no participant was excluded based on RT. Nine observations (three for each of the first three participants) were excluded, due to an error in stimulus presentation.

Raw data of all participants were transformed. Firstly, the percentage of inaccurate responses for each target word was calculated to determine which word was least accurately identified. Secondly, the percentage of inaccurate responses to each target and competitor pair was calculated to see which word-pair yielded the most inaccurate responses.

RT data was refined in two steps. The first step was to remove the incorrect responses or responses with an RT shorter than the onset of word-initial /h/+210 ms, as it takes approximately 210 ms to respond to stimulus [20]. Based on these criteria, 245 responses were excluded. Responses with too long RT were not trimmed [21] - the experiment had an inbuilt cutoff point at 2000 ms, because participants received a time-out message on screen after 2000 ms. Secondly, RT data was adjusted to two landmarks associated with the stimulus sound to accommodate the intrinsic length differences of the different vowels. The landmarks are shown in Figure 2. The first was the onset of the vowel (T1 on Figure 2), as marked by the end of the friction of /h/. The second was the offset of the vowel, (T2 on Figure 2) as marked by the closure of the /d/. Two RTs were calculated for each trial relative to the landmark vowel onset and the landmark vowel offset: RT from vowel onset is the time from the beginning of the vowel to the response and RT from vowel offset is the time from the end of the vowel to the response.

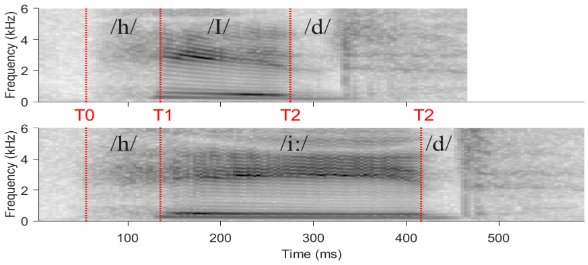


Figure 2: *Acoustic landmarks for vowel onset and offset exemplifed by the stimulus words* heed *and* hid*. T0 marks the onset of the stimulus, T1 marks the onset of the vowel and T2 marks the offset of the vowel. RT was measured from vowel onset (T1) and vowel offset (T2).*

The distribution of RT data followed Gamma distribution at both landmarks. Therefore, general-linear model with the family Gamma (GLM) was used to test RT. Gamma distribution is unable to handle the negative RT at vowel offset resulting from participants reacting before the end of the vowel. Therefore the constant 320 was added to RT from vowel offset prior to using GLM, as the lowest RT from vowel offset was −317 ms. All data analysis was conducted in R [22].

3. Results

3.1. Confusion result

The percentage of inaccurate responses per target vowel show which targets are the hardest to identify and which are the most confused vowel pairs. Table 1 shows the targets hardest to identify were /ɑe/ (4.1% of inaccurate responses), /æɔ/ (3.6%), /ɐ/ (3.4%), and /əʊ/ (3.1%). Table 1 also shows these targets also have a clear competitor, whereas /ɜː/ (the least confused vowel) was not confused with the same vowel more than once. There are also target vowels which were confused equally often with more than one competitor; in these cases all competitors were given in Table 1 and the % of errors shows the tie. However, confusion data has its limitations. Firstly, listeners are at ceiling in a binary choice task due to the high number of easy comparisons. Secondly, inaccurate answers may be a result of mispressing the buttons rather than confusing the two vowels, as 14 out of the 15 participants reported noticing they had made a mistake after pressing the answer button.

3.2. RT results

Preliminary analysis was conducted to examine the effect of target and competitor vowel on RT at vowel onset and offset. A null model without a factor and two full models with either target or competitor vowel as a factor were constructed. Target vowel had a significant effect on reaction time ($p \geq 0.0001$ at onset and offset, df=15 at both landmarks). Competitor vowel did not have an effect ($p = 0.997$ at onset and $p = 0.996$ at offset, df=15 at both landmarks). The interaction of the two factors was not significant.

To examine how target vowels affect RT, and if the 16 phonemes can be grouped according to their phonetic features, target vowels were assigned binary features following the system of [23] and using the values appropriate for AusE [17]. The features are ±high, ±low, ±front, ±back and ±long. Diphthongs were classified according to their first element, because the first element is likely to be responsible for the confusion [16, 24]. The values were coded as 0 and 1 in R. The classification of vowels is shown in Table 1.

To examine the effect of target vowel on RT relative to vowel offset and offset, a null model without a factor and five full models with each feature as a factor were constructed. At vowel onset, the features +front and +long had a significant effect, as long targets have significantly longer RT, and front targets have significantly shorter RT. At vowel offset, +low and +long had a significant effect, as long and low targets have significantly shorter RT. These results are shown in Figure 3. Testing the effects of the features of competitor vowels did not return significant results.

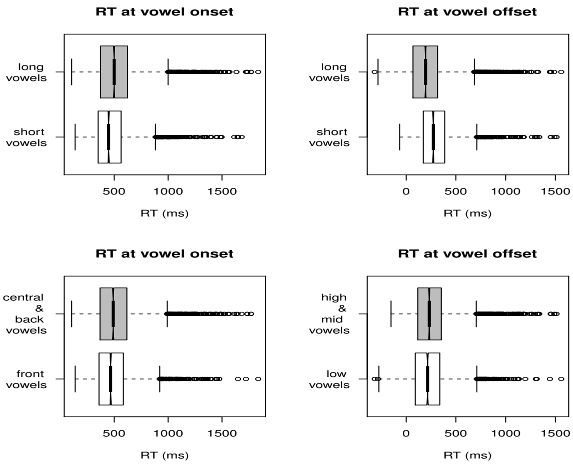


Figure 3: *RT measured from vowel onset (right panel) and from vowel offset (left panel) with the binary features ±long (upper panel), ±front (bottom left) and ±low (bottom right) as factors.*

To examine whether target vowel and competitor vowel interact five models were constructed for RT measured from vowel onset and from vowel offset. The dependent variable was RT and one feature of the target and the same feature of the competitor vowel were interacting factors. The goal was to determine if shared features of target and competitor vowel affect RT. All features except ±long interacted, showing that disambiguation is slower when target and competitor vowel share the features ±front, ±back, ±high, or ±low feature. Target and competitor length did not interact at either landmarks, showing that sharing the feature ±long does not affect disambiguation.

RT of target vowels was calculated across all competitors, and target vowels were sorted according to mean RT (from quickest to slowest). Table 2 shows that at vowel onset, short targets have short RT and long and diphthong targets have long RT. However, this is reversed at the vowel offset, when long and diphthong targets have short RT and short targets have long RT.

Table 1: *Classification of AusE vowels according to their binary features and the strongest competitor for each target according to percentage of incorrect responses and longest RT*

Target	Front	Back	High	Low	Long	Confused with	% of errors / strongest competitor	Overall accuracy (%)	Long RT
iː	+	−	+	−	+	ɪ	11	98.6	ɪ
ɪ	+	−	+	−	−	iː	11	97.8	ɐː
e	+	−	−	−	−	ɑe	4.4	98.6	iː
æ	+	−	−	+	−	æɔ	15	97.8	æɔ
ʉː	-	−	+	−	+	əʉ, ʊ	8	97.4	ʊ
ɜː	-	−	−	−	+	e, ɑe oː, ɐ, ʉː	2	99.1	ɐː
ɐː	-	−	−	+	+	æe	13	97.8	ɐ
ɐ	-	−	−	+	−	ɐː	20	96.6	ɐː
ʊ	-	+	+	−	−	oː, æɔ, oɪ	4	98.8	əʉ
oː	-	+	+	−	+	oɪ	11	98	oɪ
ɔ	-	+	−	−	−	oɪ	11	97.1	əʉ
æɪ	+	−	−	+	+	e	6	98.3	æɔ
ɑe	-	+	−	+	+	ɐː	22	95.9	ɐː
æɔ	+	−	−	+	+	æ	20	96.4	æ
əʉ	-	−	−	-	+	oɪ	11	96.9	oɪ
oɪ	-	+	+	−	+	oː, æɔ	8	97.7	æɔ

Next, two strongest competitors were selected: one with which the target was the most often confused, and a second with the longest pairwise RT. (The strongest competitor based on RT was the same at vowel onset and offset.) The strongest competitors are shown in Table 1. Table 1 also shows competitors usually share features with their targets.

Table 2: *Target vowels from shortest to longest mean RT at vowel onset and offset. Short monophthongs are in white cells, long monophthongs are in light grey cells and diphthongs are in grey cells.*

<table>
<tr><th>Time</th><th colspan="13">Target vowel</th></tr>
<tr><td>On-set</td><td>ɪ
æ
e</td><td>iː</td><td>ʊ</td><td>ɜː</td><td>oɪ</td><td>ɐ</td><td>æɪ
æɔ</td><td>ɔ</td><td>oː</td><td>ɑe</td><td>ɐː</td><td>əʉ</td><td>ʉː</td></tr>
<tr><td>Off-set</td><td>ɐː
ɜː
oː</td><td>ɑe</td><td>iː</td><td colspan="3">æɔ oɪ æɪ</td><td>æ
e</td><td>ʉː</td><td>əʉ</td><td colspan="4">ʊ ɔ ɪ ɐ</td></tr>
</table>

4. Discussion

4.1. RT results at vowel onset and offset

Short vowels have shorter RT than long vowels when RT is measured from vowel onset, and long vowels have shorter RT when RT is measured from vowel offset. The fact that short vowels have shorter RT from onset is inherent to the task, because all information on a short vowel becomes available sooner for short vowels than for long vowels. That is, listeners have access to the whole vowel sooner when the vowel is short than when it is long. The fact that the relationship between vowel length and RT is reversed at the offset shows that listeners do not need to wait until the end of the long vowel to identify it. This is further supported by the finding that RT measured from the offset is negative in 1.26% of the responses for short targets, and RT is negative in 12.5% of the responses when the target is long. There are two possible interpretations for this result. The first is that short vowels are harder to identify, and the second is that there is a minimum exposure that is required for the identification of the target vowel. The current data does not allow us to choose between these two explanations, and making this choice is beyond the scope of the present study.

Ordering target vowels from shortest to longest RT (see Table 2) points to disruptions in the general pattern. At the onset, where short targets have short RTs, long /iː/ patterns with the short vowels, and short /ɐ/ and /ɔ/ pattern with the long vowels. This may contribute to finding a significant effect of target frontness at vowel onset, as /ɔ/ is a short back vowel patterning with the long vowels, and /iː/ is a long front vowel patterning with the short vowels. Also, out of the 10 non-front vowels 7 are long, which can lead to frontness showing a significant effect. At vowel offset, low targets returned a significantly shorter RT than non-low targets; however, Table 2 shows that low vowels are spread evenly. Therefore the feature ±long seems to explain the differences in RT.

4.2. Target - strongest competitor pairs

The effect of binary features on vowel identification can be seen on the target-strongest competitor pairs. Table 3 shows that vowels confused with each other share one or more features. Also, the 5 idiosyncratic pairs share features between the target and the second element of the diphthong competitor. Vowel pairs that have the longest RT also share features.The targets /æ, æɪ, æɔ/ share both the +front and the +low features.

Table 3: *Target words (left) with their strongest competitors (right), as explained by shared features*

	Explained				Idiosyncratic
	Front	Back	High	Low	
Acc.	iː-ɪ, ɪ-iː æ-æɔ,æɪ-e, æɔ-æ	ʊ-oː,oɪ, ɔ-oɪ, oɪ-oː	iː-ɪ, ɪ-iː, ʉː-ʊ	æ-æɔ, ɐː-ɑe, ɐ-ɐː, ɑe-ɐː, æɔ-æ	e-ɑe, ʉː-əʊ, ʊ-æɔ, əʊ-ʉː, oɪ-æɔ
RT	iː-ɪ, e-iː, æ-æɔ,æɪ-æɔ, æɔ-æ	o-oɪ ʉː-ʊ, oː-oɪ	iː-ɪ, ʉː-ʊ oː-oɪ	æ-æɔ, ɐː-ɐ, ɐ-ɐː, æɪ-æɔ, ɑe-ɐː, æɔ-æ	ɪ-ɐː, ʊ-əʉ, ɔ-əʉ, oɪ-æɔ

The target-competitor pairs tend to differ in length: /iː-ɪ/, /ʉː-ʊ/, /ɐː-ɐ/, /æɔ-æ/. This confirms that AusE long-short vowel pairs are hard to disambiguate. Additionally, there are the monophthong-diphthong pairs, /æɔ-æ/, /oː-oɪ/, that share the first element and are perceptually similar. The limitation of choosing the strongest competitor for individual targets is however that the differences in RT between target-competitor pairs were small and may have been affected by lexical frequency [1].

5. Conclusions

Our results show that vowel disambiguation becomes harder when target vowel and competitor vowel share features. Disambiguation is the hardest when target and competitor only differ in length: long vowels are intrinsically hard to disambiguate from short vowels when the members of the vowel pairs have similar spectral qualities (including diphthongs that share their first element with a particular monophthong). We have identified five vowel pairs that have a high feature overlap (/iː-ɪ/, /ʉː-ʊ/, /ɐː-ɐ/, /æɔ - æ/, /oː-oɪ/) making them hard to disambiguate. Thus they are good candidates for more targeted research on the effects of phonemic similarity on vowel disambiguation.

6. References

[1] Meunier, F. and Segui, J. "Frequency Effects in Auditory Word Recognition: The Case of Suffixed Words", Journal of Memory and Language 41:327–344, 1999.

[2] Connine, M. C., Mullennix, J., Shernoff E. and Yelen, J. "Word Familiarity and Frequency in Visual and Auditory Word Recognition", Journal of Experimental Psychology: Learning, Memory, and Cognition 16(6):1084-1096, 1990.

[3] Röder, B., Demuth, L., Streb, J. and Rösler, F. "Semantic and morpho-syntactic priming in auditory word recognition in congenitally blind adults", Language and Cognitive Processes, 18(1):1-20, 2003.

[4] Vannest, J., Newport, E. L., Newman, A. J. and Bavelier, D., "Interplay between morphology and frequency in lexical access: The case of the base frequency effect", Brain Research 1373:144 – 159, 2011.

[5] Buchanan, L., Westbury, C. and Burgess, C., "Characterizing semantic space: Neighborhood effects in word recognition", Psychonomic Bulletin and Review, 8(3):531-544, 2001.

[6] Goldinger, S. D., Luce, P. A. and Pisoni, D. B., "Priming lexical neighbors of spoken words: Effects of competition and inhibition", Journal of Memory and Language, 28:501–518, 1989.

[7] Luce, P. A., Pisoni, D. B. and Goldinger, S. D., "Similarity neighborhoods of spoken words', in G. T. M. Altmann [Ed], Cognitive models of speech processing: Psycholinguistic and computational perspectives, 122–147, MIT Press 1990.

[8] Cluff, M. S. and Luce, P. A., "Similarity neighborhoods of spoken two-syllable words: Retroactive effects on multiple activation.", Journal of Experimental Psychology: Human Perception and Performance, 16:551–563 1990.

[9] Ziegler, J. C., Ferrand, L. and Montant, M., "Visual phonology: The effects of orthographic consistency on different auditory word recognition tasks", Memory and Cognition 32(5):732–741, 2004.

[10] Assmann, P. F., Nearey, M. T. and Hogan, T. J., "Vowel identification: Orthographic, perceptual, and acoustic aspects", The Journal of the Acoustic Society of America 71:975-989, 1982.

[11] McMurray, B., Tanenhaus, M. K., Aslin, R. N., and Spivey, M. J., "Probabilistic Constraint Satisfaction at the Lexical /Phonetic Interface: Evidence for Gradient Effects of Within-Category VOT on Lexical Access" Journal of Psycholinguistic Research 32:77-97, 2003.

[12] Grosjean, F., "Spoken word recognition processes and the gating paradigm", Perception and Psychophysics 28:267-283, 1980.

[13] McQueen, J. M. and Viebahn, M. C., "Tracking recognition of spoken words by tracking looks to printed words", The Quarterly Journal of Experimental Psychology 60:661-671, 2007.

[14] Allopenna, P. D., Magnuson, J. S., and Tanenhaus, M. K., "Tracking the Time Course of Spoken Word Recognition Using Eye Movements: Evidence for Continuous Mapping Models", Journal of Memory and Language 38:419439, 1998

[15] Taft, M., "Lexical access codes in visual and auditory word recognition", Language and Cognitive Processes, 1(4):297–308, 1986.

[16] Cox, F. "Vowel Change in Australian English", Phonetica 56:1–27, 1999.

[17] Cox, F., Australian English Pronunciation and transcription, Cambridge University Press, 2012.

[18] Cutler, A., Native Listening: Language Experience and the Recognition of Spoken Words, MIT Press, 2012.

[19] Psychology Software Tools, I. E-Prime 2.0 2012.

[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 in Human Neuroscience 9(131):1-12, 2015.

[21] Ratcliff, R., "Methods for dealing with reaction time outliers", Psychological Bulletin 114(3):510-532, 1993.

[22] R version 3.2.3 2015.

[23] Chomsky, N. and Halle, M., The sound pattern of English, MIT Press, 1968.

[24] Nearey, T. M. and Assmann, P. F., "Modeling the role of inherent spectral change in vowel identification", The Journal of the Acoustic Society of America 80:1287-1308, 1986.

Regional priming in Australian English KIT, DRESS and TRAP vowels

Michael Walker, Anita Szakay, Felicity Cox

Department of Linguistics, Macquarie University, Australia

michael.walker@mq.edu.au, anita.szakay@mq.edu.au, felicity.cox@mq.edu.au

Abstract

40 speakers of AusE participated in a pseudo-replication of Hay and Drager's stuffed toy study [1]. Participants heard spoken AusE sentences with a phrase-final target word containing either a KIT, DRESS or TRAP vowel [2] followed by a six-step continuum of isolated synthesised vowels, from NZE-like to exaggerated AusE-like. Participants selected the token from the continuum they believed best matched the realisation of the vowel in the target word. Participants were exposed to stuffed toy kiwis, koalas, or neither (control). Contrary to results reported in [1], participants' responses were not influenced by the presence of the toys.

Index Terms: speech perception, vowel perception, sociophonetics, Australian English, New Zealand English

1. Introduction

Reference to a particular country or dialect has been shown to influence listeners' performance in perceptual tasks [1], [3], [4]. Listeners are thought to be highly sensitive to the sociophonetic consequences of a speaker's social characteristics, such as nationality, and may anticipate certain features in speech, resulting in predictable perceptual biases [4]. These findings support the idea that indexical, as well as linguistic information plays a role in speech perception.

Niedzielski [3] showed that the expected dialect of a speaker could influence the categorisation of socially salient dialectal variation. Canadians and Detroiters both typically produce a raised MOUTH vowel, consistent with vowels influenced by Canadian Raising (CR), yet Detroiters do not recognise the raising in their own speech. Participants from Detroit heard recorded speech from a single Detroit speaker but were told the speaker was from either Canada or Detroit, with response sheets also labelled *Canadian* or *Michigan*. Phrases with words containing a target MOUTH vowel were played, followed by a continuum of six synthesised tokens ranging from raised to unraised variants of the target vowel. Participants selected the token they believed was most like the vowel in the target word. Participants in the Canadian condition selected on average a more raised variant than those in the Michigan condition. [3] argues that because Detroiters only associate CR with Canadian English they would not expect to hear raised variants in a speaker of their own dialect.

Hay, Nolan and Drager [4] replicated the procedure outlined in [3] by testing New Zealanders' perception of KIT, DRESS and TRAP vowels produced by a speaker of New Zealand English (NZE). Participant answer sheets were labelled with either *Australia* or *New Zealand*. Continua consisted of six tokens from Australian English (AusE)-like to NZE-like. Perception of vowels was shown to shift in the direction of well-recognised differences between AusE and NZE, consistent with the priming condition. However, this effect was limited to female participants. KIT vowels, being the most salient difference between AusE and NZE, showed the highest variability congruence with the prime. A lesser effect was observed in TRAP vowels, and DRESS vowels showed no influence of the prime. In a post-task questionnaire, all participants, except one, identified the speaker as a NZE speaker. It was suggested that activation of the concept of Australia, rather than participants believing the speaker was Australian, was enough to shift perception of vowels towards values associated with AusE [4].

Hay and Drager [1] proposed that any culturally significant token might produce a similar effect. In [1], participants completed the same task as [4], using identical auditory stimuli. Stuffed toy koalas and kangaroos were used to activate 'Australia' resulting in NZE-speaking participants selecting more AusE-like tokens when compared with a group exposed to stuffed toy kiwis. However, the effect was again limited to female participants and KIT vowels.

In another replication of [3], Lawrence [5] found no evidence for regional labels ('Sheffield, Northern England' or 'London, Southern England') influencing responses in BATH and STRUT vowels. Despite BATH and STRUT being highly salient regional markers, no priming effect was found. This suggests the priming effect observed in [1], [3], [4] may be highly contextually specific and not generalisable.

The present study is a pseudo-replication of the experiment outlined in [1], [3], [4]. It seeks to determine whether a shift in perception as a consequence of the stuffed toy prime, as reported in [1], would occur in an AusE-speaking sample. This study also aims to address a number of design and procedural concerns in [1], [3], [4] relating to control of stimuli, continua construction and the lack of a control condition. Consistent with [1], [4] the experiment focuses on KIT, DRESS and TRAP vowels in AusE and NZE.

The NZE KIT is relatively centralised and typically realised as [ɘ] [6] while in AusE it is realised as [ɪ] [7]. The difference in KIT between the two dialects is said to be the most salient and well known to laypeople [4] as exemplified in the stereotyped "fush and chups" (NZE) vs "feesh and cheeps" (AusE). The NZE DRESS is typically a more raised variant of /e/ than the AusE counterpart [8]. TRAP is typically realised as [ɛ] in NZE and [æ] in AusE [8**]**.

Hay et al. [1], [4] draw on an exemplar model of speech perception to account for the priming effect shown in [1], [3], [4]. The model proposes that linguistic information is represented in memory as phonetically detailed exemplars. Exemplars form categories which represent a class of equivalent perceptual experiences [10]. Social information about the speaker is also retained. Categorisation of new input involves comparison to existing exemplars [10] and is said to be biased towards frequently activated acoustic and social categories [4]. Accordingly, by activating 'New Zealand' with stuffed toy kiwis, Australian listeners would be biased towards categorising speech input as NZE, provided the exemplars associated with NZE are robust.

Following the results presented in [1], [3], [4] and assuming an exemplar model of speech perception, we expect that, for those AusE-speaking participants exposed to the kiwi, the perception of target vowels would shift towards more NZE-like vowel qualities. Consistent with [1], [4] this shift should be present in words containing KIT vowels but may not be with DRESS and TRAP vowels. However, according to an exemplar model, associations between social and linguistic categories emerge only after frequent activation. For this reason, we predict that the priming effect will be limited to those participants who indicate some level of exposure to New Zealand and NZE.

2. Methods

2.1. Participants

40 female speakers of AusE participated in the perception task. 3 males also participated but their data will not be discussed here. All were born and educated in Australia and were undergraduate students at Macquarie University in Sydney receiving course credit for their time. Ages ranged from 18-27 with a mean age of 19.5.

2.2. Stimuli and materials

2.2.1. Target vowels and target words

KIT, DRESS and TRAP target vowels were represented in /CVt/ or /CCVt/ monosyllabic words. 10 different words were used for each of the three target vowels. Due to lexical restrictions, some complex onsets were included (*grit*, *skit*, *slat*, *Brett*, *threat*). The set of target words represent more controlled stimuli than those used in [1], [4], which used multiple coda consonants.

2.2.2. Sentences

Target words were each embedded in 30 unique sentences. Each sentence contained the target word in phrase-final position thus participants would not be exposed to any additional vowels between the target and continuum tokens. In addition, consistency in target word position allowed for straightforward target identification by participants. Unlike the stimuli used in [1], [4], there were no other examples of the target vowels in stressed position in any sentence. This minimised any additional priming effect by reducing overt identifiers of AusE (when contrasted with NZE**)**. Example sentences are shown:

1. The new movie was a huge summer **hit**
2. She is studying to become a **vet**
3. The bug looked like it was a **gnat**

A 19-year-old male monolingual speaker of Standard AusE from Sydney read the sentences. Stimuli were recorded in a soundproof room with an AKG C535 condenser microphone and a PreSonus StudioLive 16.4.2 digital mixer using Pro Tools 11.3.1 at a 48kHz sampling rate.

2.2.3. Continua

Continua consisted of six synthesised vowel tokens representing equal steps from NZE-like (token 1) to exaggerated AusE-like (token 6). Characteristics of the speaker's spoken target vowel from each sentence (i.e. F_1 and F_2) were used to synthesise token 4 in Praat [11]. F_1 and F_2 values of the synthesised vowels were then manipulated to create the additional five tokens (see Table 1 for examples).

Table 1. *Formant values for 'hit', 'vet' and 'gnat' continuum tokens.*

Token	hit		vet		gnat	
	F1	F2	F1	F2	F1	F2
1	484	1627	380	2078	655	1995
2	445	1781	466	1987	722	1852
3	407	1948	560	1900	792	1718
4	370	2130	660	1817	866	1594
5	334	2330	769	1737	944	1480
6	300	2551	887	1661	1027	1371

Step intervals between continua tokens were calculated in bark steps [12]. For KIT vowels, intervals were 0.35 bark steps for F_1 and 0.6 bark steps for F_2. For DRESS the intervals were 0.8 (F_1) and 0.3 (F_2). For TRAP, 0.5 (F_1) and 0.5 (F_2). Step intervals were calculated such that F_1 and F_2 values for token 1 were within 2 standard deviations of mean NZE values published in [13]. Mean AusE values provided by [14] were used as a reference for tokens 4-6. This differs from the continua used in [1], [4], which were based on a single vowel produced by the speaker in a /hVd/ context. This meant that in [1], [4] continua tokens were consistent but token 4 would not align exactly with the speaker's vowel in each target word. We elected to create a unique continuum for each target word where the F_1 and F_2 values of token 4 matched those of the vowel produced by the speaker in that sentence. Synthesised tokens were 180ms in duration. Initial F_0 was 160Hz (the mean initial F_0 for all tokens) with an F_0 slope of -1 octaves per second applied to each of the synthesised vowels. Figure 1 shows spectrograms of the synthesised continuum tokens for *hit*.

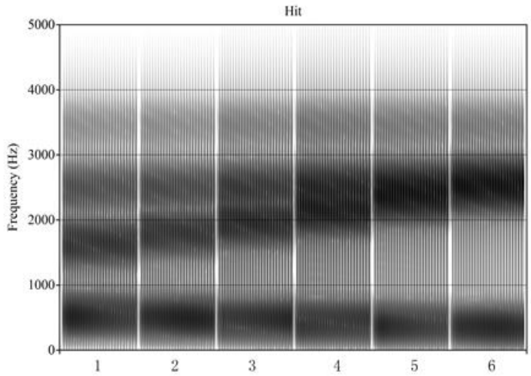


Figure 1. *Spectrogram of continuum tokens for 'hit' from most NZE-like to exaggerated AusE.*

2.3. Procedure

2.3.1. Priming

Participants were exposed to one of three conditions: Australian (n = 13), New Zealand (n = 14) and a control (n = 13). Similar to [1], priming for the conditions was cued by the presence of stuffed toy koalas (Australian) or kiwis (New Zealand) with the control exposed to no toys. To introduce the prime, the experimenter 'found' the headphones required for the task in a drawer under the toys. The experimenter mentioned the toys were being used in another experiment and placed them on the table beside the computer being used for the task. The intent was to draw attention to the prime without it being obvious that it was related to the experiment. The toys remained in the participant's line of sight for the duration of the experiment. This was similar to the procedure used in [1]. A control was not used in [1], [3-5], however, it was deemed important to test the perception task in an un-primed context.

Each participant completed the task individually in a soundproofed room, which contained no other regional primes. The only difference between the conditions was the presence of the prime itself. All participants interacted with the same experimenter – a 29-year-old male, born in New Zealand who had moved to Australia at age 10 – so any potential priming effect from the experimenter is consistent for all participants.

2.3.2. Perception task

Participants heard each sentence while it was simultaneously presented on-screen with the target word bold and underlined. Sentences remained visible for the duration of the recorded stimuli plus an additional 2000ms. Following the sentence, continuum tokens were played with the corresponding number for each token visible for 1000ms. After hearing all six continuum tokens, participants were prompted to make their selection by pressing the corresponding number key. The task did not continue until a selection was made. Each sentence was presented once with the continuum tokens in order from token 1 to token 6, and once with the continuum tokens in reverse order. Presentation of the synthesised tokens with a visual numeric label differs slightly from [1], [4] who used spoken numbers preceding each token to label the steps. Visual labelling was preferred to reduce any additional priming effect, particularly from the word *six*, which contains a KIT vowel. Prior to the main task, participants completed three familiarisation questions using non-target vowels with stimuli read by a different male speaker.

The task was divided into two blocks. Each block contained all 30 sentences once, half with their continua presented in the original order and half with their continua presented in reversed order. This meant that the synthesised version of the speaker's vowel, as well as highly centralised or peripheral variants, would not always be in the same position, discouraging any response patterning. Sentence order was randomised within each block with each sentence heard once per block and twice overall. The perception task was presented on a Sony Viao laptop using E-Prime 2.0 [15]. All participants used Sennheiser HD 461i closed over-ear headphones and could adjust volume to a comfortable level.

Following the perception task, participants completed two questionnaires modelled on those used in [4]. The first concerned the participant's impressions of the speaker's age, nationality, occupation and education level. The second was designed to assess participant's level of exposure to New Zealand and NZE.

3. Results

Table 2 shows mean selection scores for KIT, DRESS and TRAP vowels in the three conditions, as well as the mean score across all conditions. On average, participants selected tokens that represent more phonetically raised variants of KIT, DRESS and TRAP than the vowel produced by the speaker. This explains why the mean token selection for KIT is so different to the mean selections for both DRESS and TRAP. For KIT vowels, higher token numbers represent raised and fronted variants (more AusE-like) however, for the DRESS and TRAP vowels, lower token numbers represent raised and fronted variants (more NZE-like). Figure 2 shows the distribution of responses for KIT vowels in the three conditions.

Table 2. *Mean token selection.*

	Aus	NZ	Control	All
KIT	5.21	5.23	5.02	5.16
DRESS	3.13	3.10	3.06	3.10
TRAP	3.40	3.55	3.38	3.45

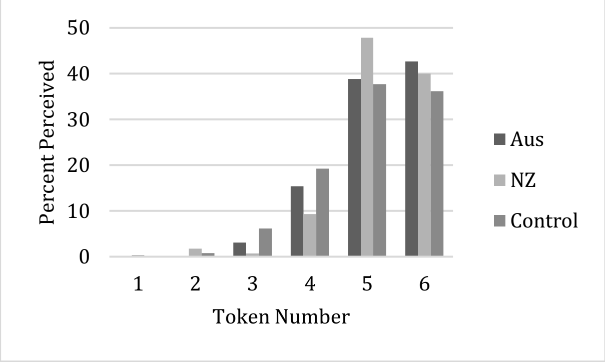


Figure 2. *Token selections for KIT vowels. 1 is NZE-like and 6 is exaggerated AusE.*

We fit a linear mixed effects model [16] to the data using the lme4 library in R [17], [18] with random intercepts for speaker and word. Fixed effects included experimental condition (Aus, NZ and control) and vowel class (KIT, DRESS and TRAP). Exposure to NZE (no frequent exposure, frequent exposure) was originally included in the model, however, it was not found to be a predictor of participants' responses, therefore it was removed. Significance was calculated using Satterthwaite's [19] approximations for the degrees of freedom using the lmerTest library [20]. As shown in Table 3, no significant difference was found between responses in the control condition and either primed condition. The model was also run with Australia as the default level of condition. This showed that the Australia and New Zealand conditions were also not significantly different from each other (p=0.6). Contrary to findings in [1], the presence of the stuffed toy koalas or kiwis was not shown to have a significant influence on responses in the perception task. Table 3 also shows that mean selections for KIT vowels differed significantly from TRAP. Selections also varied significantly between KIT and DRESS (p<0.001).

Table 3. *Fixed effects with control as default condition and TRAP as default vowel*

	Estimate	Std. error	df	t-Value	p-Value
(Intercept)	3.19	0.10	55.6	31.52	< 0.001
Condition = NZ	0.14	0.09	37	1.60	0.117
Condition = Aus	0.10	0.09	37	1.09	0.281
Vowel = DRESS	-0.35	0.11	27	-3.10	0.004
Vowel = KIT	1.71	0.11	27	15.23	< 0.001

Although the model did not show significant differences in token selection between conditions, participants did respond differently in a free-choice task regarding the speaker's nationality. Participants were less likely to identify the speaker as 'Australian' in the New Zealand condition (71%) than in the Australian or control conditions (92%).

4. Discussion

The priming effect shown in [1] was not replicated in an Australian context. Participants exposed to stuffed toy kiwis did not perform differently in the matching task when compared with those exposed to toy koalas or those in the control condition. If the effect relies on a strong association between the toy, its elicited dialect, and relevant sociophonetic variants, this result may simply suggest that the toy kiwis were not culturally significant enough to activate 'New Zealand' for the Australian participants. In other words, there is no way of knowing whether those participants in the New Zealand condition recognised the toys as kiwis and, in turn, associated them with New Zealand.

Indeed, questionnaire responses indicate that 11 participants had never been to New Zealand, didn't speak with, or know, any New Zealanders and couldn't name any New Zealand media. While these individuals may be generally aware of NZE and how it differs from AusE, NZE is a foreign dialect in Australia so phonetic sensitivity should not be assumed. If this is the case, then the lack of a result supporting [1] is not surprising. It would have been useful to have participants in the New Zealand condition tested on their ability to identify AusE and NZE in a post-experiment task. If participants completing the task could be shown to identify a NZE speaker with a level of accuracy similar to that reported in [9] yet still not show a priming effect, this would strengthen the null result.

We predicted that, in the New Zealand condition, participants with some level of exposure to NZE and New Zealanders would select more NZE-like vowels, showing influence of the prime. More than half (n = 26) of participants reported speaking with New Zealanders on a regular basis or having travelled to New Zealand. Six participants had a parent, partner or close friends from New Zealand. However, responses in the matching task from these participants did not show the predicted effect suggesting that our null result cannot be attributed to a lack of exposure alone. Either the influence of social information is more limited than suggested in [1], [3], [4] or the influence exists but is highly contextually specific. Further, it may be that even with frequent exposure to NZE, an individual may not have sufficient phonetic sensitivity to complete the matching task as predicted.

The lack of a priming effect in the present study does not necessarily contradict Hay et al.'s [1], [4] argument for exemplar-based speech categorisation. It may be that, even for those Australians with frequent exposure to NZE, NZE exemplars or categories are not activated enough, if at all, by the kiwi to compete with the resting activation level of AusE exemplars. A more overt priming condition might be required to produce a shift in Australians' responses towards more NZE-like variants. A possible priming effect from the kiwis was observed as it reduced the proportion of participants who recognised the speaker as Australian. However, responses did not correlate with the prime, as only one participant answered 'New Zealand' when identifying the speaker's nationality.

5. Conclusion

The stuffed toy priming effect observed in [1] was not replicated in an Australian context, even for those participants who had frequent contact with NZE. Further experimentation in this paradigm may be able to establish whether the results reported in [1], [3], [4] are limited to highly contextually specific situations or whether the participants tested in the present experiment and [5] simply did not have sufficient exposure to the primed dialect.

6. References

[1] J. Hay and K. Drager, "Stuffed toys and speech perception," *Linguistics*, vol. 48, no. 4, pp. 865–892, Jul. 2010.

[2] J. C. Wells, *Accents of English.* Vol. 1. Cambridge University Press, 1982.

[3] N. Niedzielski, "The Effect of Social Information on the Perception of Sociolinguistic Variables," *Journal of Language and Social Psychology*, vol. 18, no. 1, pp. 62–85, Mar. 1999.

[4] J. Hay, A. Nolan, and K. Drager, "From fush to feesh: Exemplar priming in speech perception," *Linguistic Review*, vol. 23, no. 3, pp. 351–379, Sep. 2006.

[5] D. Lawrence, "Limited evidence for social priming in the perception of the bath and strut vowels," in *International Congress of Phonetic Sciences*, Glasgow, 2015.

[6] L. Bauer, P. Warren, D. Bardsley, M. Kennedy, and G. Major, "New Zealand English," *Journal of the International Phonetic Association*, vol. 37, no. 1, pp. 97–102, Apr. 2007.

[7] J. Harrington, F. Cox, and Z. Evans, "An acoustic phonetic study of broad, general, and cultivated Australian English vowels," *Australian Journal of Linguistics*, vol. 17, no. 2, pp. 155–184, Sep. 1997.

[8] C. I. Watson, J. Harrington, and Z. Evans, "An acoustic comparison between New Zealand and Australian English vowels," *Australian Journal of Linguistics*, vol. 18, no. 2, pp. 185–207, Oct. 1998.

[9] I. Ludwig, "Identification of New Zealand English and Australian English based on stereotypical accent markers," M.A. thesis, Univ. of Canterbury, New Zealand, 2007.

[10] J. Pierrehumbert, "Exemplar dynamics: Word frequency, lenition and contrast. Frequency and the emergence of linguistic structure," in *Frequency effects and the emergence of linguistic structure*, J. Bybee and P. Hopper, Eds. Amsterdam: John Benjamins, 2001, pp. 137-57.

[11] P. Boersma and D. Weenink, *Praat: Doing phonetics by computer* (Version 5.4) [Computer program], 2009.

[12] E. Zwicker, "Subdivision of the Audible Frequency Range into Critical Bands (Frequenzgruppen)," *The Journal of the Acoustical Society of America*, vol. 33, no. 2, pp. 248–248, Feb. 1961.

[13] A. Easton and L. Bauer, "An Acoustic Study of the Vowels of New Zealand English," *Australian Journal of Linguistics*, vol. 20, no. 2, pp. 93–117, Oct. 2000.

[14] F. Cox, "Formant Frequencies and Durations for /hVd/ vowels from AusTalk," unpublished.

[15] Psychology Software Tools, Inc, *E-Prime 2.0* (Version 2.0) [Computer program], 2012.

[16] H. Baayen, *Analyzing Linguistic Data; A Practical Introduction to Statistics using R*. Cambridge University Press, 2008.

[17] D. Bates, M. Maechler, and B. Bolker, *Linear mixed-effects models using S4 classes* (R-Version 0.999375–41), 2011.

[18] R Core Team, *R: A language and environment for statistical computing* (Version 3.3.1) [Computer program], 2016.

[19] F. E. Satterthwaite, "An Approximate Distribution of Estimates of Variance Components," *Biometrics Bulletin*, vol. 2, no. 6, pp. 110–114, 1946.

[20] A. Kuznetsova, P. B. Brockhoff, and R. H. B. Christensen, *lmerTest: tests for random and fixed effects for linear mixed effect models (lmer objects of lme4 package)* (R-Version: 1.1-0), 2013.

The Relative Contributions of Duration and Amplitude to the Perception of Japanese-accented English as a Function of L2 Experience

Saya Kawase, Jeesun Kim, Chris Davis

The MARCS Institute, Western Sydney University, Sydney, NSW, Australia

s.kawase@westernsydney.edu.au

Abstract

This study investigated the effect of duration and amplitude of speech produced by non-native (L2) talkers on foreign accent and comprehensibility ratings. We selected two Japanese L2 talkers with different experience in producing L2 English and then transplanted their speech duration and amplitude values onto English native speech. Native English listeners rated the transplanted sentences as well as untransplanted ones (control). Transplanted sentences were rated as more accented and more difficult to understand than control sentences. The degree of L2 experience only made a difference for the perceived degree of foreign accent ratings when duration was transplanted.

Index Terms: foreign-accented speech perception, speech rhythm, second language speech

1. Introduction

Speech produced by second language (L2) talkers is often perceived as having a ‘foreign accent’. A foreign accent typically occurs due to the influence of a talker’s first language (L1), namely differences between L1 and L2 at segmental and/or supra-segmental levels. For example, in terms of segments, an accented L2 talker may miss, insert or substitute phones. At the supra-segmental level, an L2 talker’s speech rhythm, intonation or prosody may be different.

Foreign-accented speech is typically less intelligible and/or comprehensible than native speech and requires more listening effort [1]. Interestingly, the degree of perceived accent does not necessarily correspond to the degree of recognition difficulties. For example, listeners may be able to correctly identify words spoken by L2 talkers and yet perceive these as ‘heavily accented’ [2]. Likewise, the perceived degree of foreign accent does not always match with ratings of listener effort required in understanding the speech [3]. For example, in Derwing and Munro [3], native English listeners were presented with spoken sentences produced by foreign-accented talkers from four language background (Cantonese, Japanese, Polish, and Spanish) and were asked first to transcribe the utterances, then to rate their degree of foreign accent (from “no accent” to “extremely strong accent”) and after a short break to rate how easily they could understand them (from “extremely easy to understand” to “extremely difficult or impossible to understand”). The results showed that the accent ratings tended to be more extreme than the ease of understanding ratings, which were harsher than intelligibility scores (based on transcription errors). This indicates that the salience of an accent in foreign speech may not be commensurate with the effort required to understand it or its intelligibility.

In evaluating how foreign accent affects speech perception (ease and intelligibility) it is important to consider both the segmental and suprasegmental levels. Even though previous results show that the intelligibility of L2 speech and the extent that a foreign accent is perceived can be explained by differences at both segmental [e.g., 4-7] and prosodic levels [e.g., 8-10], the bulk of research has focused on the possible effects from segmental differences. For example, extensive studies have shown the difficulties in producing non-native phonemes such as English /l-r/ distinction by Japanese learners of English [e.g., 4]. Research in L2 vowel production also showed the influence of different vowel inventory size between L1 and L2 [6]. L2 learners face difficulty in producing non-native vowel contrasts especially among the learners with a smaller vowel inventory (e.g., five in Japanese) when producing vowels with a larger vowel inventory (e.g., Australian English) [7].

Relatively fewer studies have examined how non-native suprasegmental differences contribute to foreign accent. Such difference can be glossed in terms of L2 speakers having a non-native prosody. Speech prosody refers to features of speech such as pitch, stress, and duration [11], and these features also differ cross-linguistically. For example, differences in how the rhythms of languages pattern, have led to the idea of different rhythm classes. This typology groups languages into so called stress-timed (e.g., English, Dutch), syllable-timed (e.g., Spanish, French) and mora-timed (e.g., Japanese) classes [12, 13].

By and large, it has been found that the characteristics of L2 rhythm are influenced by the L1 rhythm type [e.g., 8, 9]. These studies have used duration-based measures to characterize L2 rhythm, however, recent studies suggest that patterns in speech amplitude also provide an index of rhythm. For example, a rhythm metric has recently been developed on the basis of changes in speech amplitude. This metric uses the amplitude modulation (AM) structure of the speech envelope and was derived from the Spectral Amplitude Modulation Phase Hierarchy (S-AMPH) model [14]. This rhythm metric is different from previous durational statistic-based approaches since it only considers slow changes (modulations) in the amplitude of acoustic signals. Our preliminary analysis using this model suggests differences in AM between L1 English and L2 English (produced by Japanese talkers). While further studies are required, different amplitude patterns between L1 and L2 talkers likely exist and so will contribute to foreign-accent. Taking the above into account, in the current study we investigated the extent to which L2 speech rhythm affects the ease of perceiving speech and the degree of foreign accent and we evaluated L2 rhythm using both duration and amplitude measures.

Before detailing the experimental method, we consider one other prominent factor that has been found to affect the perceived degree of foreign accent, L2 talker’s experience. Models of learning L2 speech (e.g., Speech Learning Model (SLM) [15, 16] and Perceptual Assimilation Model of Second

Language Speech Learning (PAM-L2) [17]) have postulated that L2 learners will improve in their perception of non-native contrasts with experience of the L2 speaking environment. Consistent with this, research comparing Japanese learners of L2 (English) vowels with more than 12 months of residence in an English speaking country to those with 6 months or less, has shown significant improvement in the former compared to the latter [7] (although this was not the case for difficult contrasts, e.g., tense-lax vowels). Given these models and studies have focused on segmental learning, the extent to which L2 prosody can be acquired is still largely unknown. Thus, it is important to investigate how L2 prosody is produced and perceived as a function of L2 experience.

One study has investigated the influence of L2 experience on speech intelligibility and accentedness with English-French bilinguals [18]. In this study, the speech of L2 talkers that had different experience levels (inexperienced and experienced French talkers of English) was used as the basis for a prosody 'transplantation' study. In this approach, certain prosodic features (segment duration, amplitude and F0 contour) from the L2 English utterances were transplanted onto ones spoken by native English speakers. Then these sentences were tested for intelligibility and rated for foreign accent. The results showed that transplanting features of the inexperienced French speaker's English prosody onto English native segments decreased intelligibility and increased accentedness ratings.

The current study followed up [18] by investigating the effect of non-native speech rhythm on the perceived degree of foreign accent with Japanese-accented English as a function of L2 experience. In order to assess the influence L2 rhythm on the perception of foreign-accented speech, we used a Pitch-synchronous overlap-add (PSOLA) procedure [21] that allowed the L2 duration and amplitude produced of an experienced and an inexperienced Japanese talker to be transplanted onto the segments produced by a native Australian English talker. The transplanted speech was then presented to native Australian English listeners who were asked to rate the degree of foreign accent (accent rating task) and how easily they could understand (ease of understanding; also known as comprehensibility [1-3]) that speech using a 9-point scale. It was expected that perceived degree of foreign accent and comprehensibility would be affected by L2 talker's experience. Specifically, higher accentedness and reduced ease of understanding were predicted for the inexperienced Japanese-English L2 speech compared to the more experienced ones.

2. Methods

2.1. Participants

Forty five native Australian English listeners (34 females, 11 males; M_{age} = 22.5 years) participated in this study. They were recruited from the Western Sydney University using the university's research participation system. All of the participants reported normal hearing. From a questionnaire we ascertained that none of the participants were familiar with Japanese-accented English.

2.2. Stimuli

2.2.1. Materials/Talkers

The stimuli consisted of 56 IEEE Harvard Sentences produced by two Japanese talkers and one Australian English talker (all females; M_{age} = 24.0 years) who resided in Sydney, Australia. The Japanese talker who was considered to be "inexperienced" in English had fairly recently arrived in Sydney (length of residence (LOR) = 4.5 months) and the "experienced" Japanese talker had been in Sydney more than a year (LOR = 12.5 months). According to our separate foreign accent rating study by native Australian English listeners (n = 15), the inexperienced Japanese talker's English was perceived as "strongly foreign-accented" (8.1 out of 9) and the experienced Japanese talker's English was perceived as "mildly foreign-accented" (4.7 out of 9). The monolingual Australian English talker was born and raised in Sydney, and was recruited at Western Sydney University. All talkers reported no history of speech, vision or hearing problems.

2.2.2. Stimulus editing (PSOLA)

The recording was made using an externally connected lapel microphone, (an AT4033a audio-technica microphone) in 44.1 kHz, 16-bit mono. Following the recording, all the stimulus sentences were segmented into phonemes. A Pitch-synchronous overlap-add (PSOLA) procedure [21] was then used to impose L2 durations and amplitude onto the native segments. Table 1 describes the details of how segments and prosody were combined: (1) native English segments with inexperienced Japanese amplitude (NE_NJI:I); (2) native English segments with inexperienced Japanese durations (NE_NJI:D); (3) native English segments with experienced Japanese amplitude (NE_NJE:I); and (4) native English segments with experienced Japanese durations (NE_NJE:D). In addition to the four manipulated conditions, another condition, native English (no manipulation), was also prepared.

Table 1. *Illustration of stimulus manipulations*

Segment	Transferred Prosody	Manipulation Types	Codes
Native English	NA	NA	NE
	Inexperienced Japanese	Amplitude	NE_NJI:I
		Duration	NE_NJI:D
	Experienced Japanese	Amplitude	NE_NJE:I
		Duration	NE_NJE:D

2.3. Procedure

2.3.1. Accent Rating Task

The participants were tested individually in a sound-treated booth. All of the participants performed an accent rating task [cf. 1-3]. They were required to listen to the English sentences while paying careful attention to the accent. They were then asked to rate the degree of accent on a 9-point scale (from 1: no foreign accent at all to 9: very strong foreign accent). They were encouraged to use the full rating scales for the judgments. Both rating tasks started with practice trials (n = 4) to become familiar with the rating scales.

2.3.2. Comprehensibility Rating Task

After completing the accent rating task, a comprehensibility (ease of understanding) rating task was conducted. In this task, the participants were instructed to listen to each sentence and were asked to rate how easily they could understand it using a 9-point scale (from 1: easy to understand to 9: very difficult to understand). Here, the intention was to measure listener effort in the perception of foreign-accented speech [cf. 2]. Note, the same stimulus materials were used as the accent rating task.

3. Results

Separate analyses for accent rating and comprehensibility rating data were conducted with one-way ANOVAs with lm in R 3.2.1. to predict accent rating score (3.1) and comprehensibility rating score (3.2) based on different prosodic manipulation types.

3.1. Accent Rating

Figure 1 shows mean accent rating scores for each prosody manipulation condition. Across the participants, the mean accent rating score in Native English (both segments and prosody) was perceived lower (less accented) than the rest of manipulated conditions, NE_NJE:D (β = 0.75, SE = 0.08, t = 9.08, $p < .0001$), NE_NJI:D (β = 0.99, SE = 0.08, t = 11.962, $p < .0001$), NE_NJE:I (β = 0.30, SE = 0.08, t = 3.59, $p < .0001$), NE_NJI:D (β = 0.26, SE = 0.08, t = 3.116, $p < .001$). Furthermore, we ran multiple comparisons between stimulus manipulation types (duration and amplitude) and L2 experience types (experienced vs. inexperienced) using the glht function in R package multcomp [22].

The results showed stronger foreign-accented durational influences compared to the amplitude ones on perceived degree of foreign accent in both experienced Japanese (β = -0.46, SE = 0.10, t = -4.76, $p < .0001$) and inexperienced Japanese (β = -0.73, SE = 0.10, t = -7.76, $p < .0001$). In addition, we found that the experienced Japanese non-native prosody was perceived less accented compared to the inexperienced Japanese only in the duration manipulation (β = 0.24, SE = 0.10, t = 2.50, $p < .05$), but not in the amplitude one ($p > .05$).

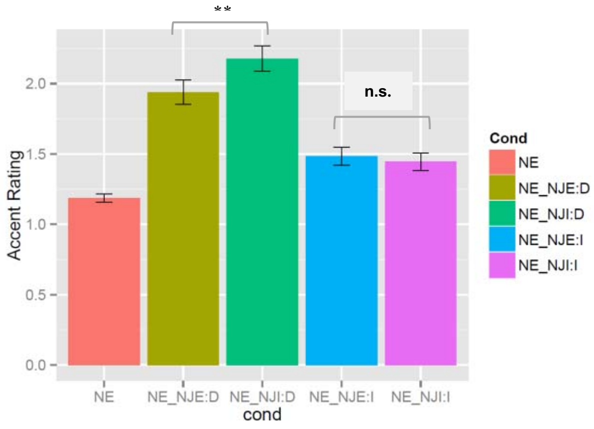


Figure 1: Mean accent rating scores for each prosody manipulation conditions. (NE: Native English, NE_NJE:D: Native English segments with experienced Japanese durations, NE_NJI:D: Native English segments with inexperienced Japanese durations, NE_NJE:I: Native English segments with experienced Japanese intensity, and NE_NJI:I: Native English segments with inexperienced Japanese intensity). Error bars indicate +/- one standard error.

3.2. Comprehensibility Rating

Figure 2 shows mean comprehensibility rating scores for each prosody manipulation condition. The mean comprehensibility rating score in Native English (both segments and prosody) was lower (i.e., easier to be understood) compared to the other manipulated conditions, NE_NJE:D (β = 0.67, SE = 0.76, t = 8.78, $p < .0001$), NE_NJI:D (β = 0.84, SE = 0.76, t = 10.97, $p < .0001$), NE_NJE:I (β = 0.35, SE = 0.76, t = 4.56, $p < .0001$), NE_NJI:D (β = 0.31, SE = 0.08, t = 4.11, $p < .001$). As with the previous analyses, we ran multiple comparisons between stimulus manipulation types (duration and amplitude) and L2 experience types (experienced vs. inexperienced).

Similar to the accent rating data, there were stronger durational influences compared to the amplitude ones on comprehensibility in both experienced Japanese (β = -0.32, SE = 0.08, t = -3.6, $p < .0001$) and inexperienced Japanese (β = -0.52, SE = 0.08, t = -5.93, $p < .0001$). However, unlike the accent rating scores, we found no significant difference in comprehensibility scores between the experienced and the inexperienced Japanese stimuli in both duration and amplitude ($p > .05$).

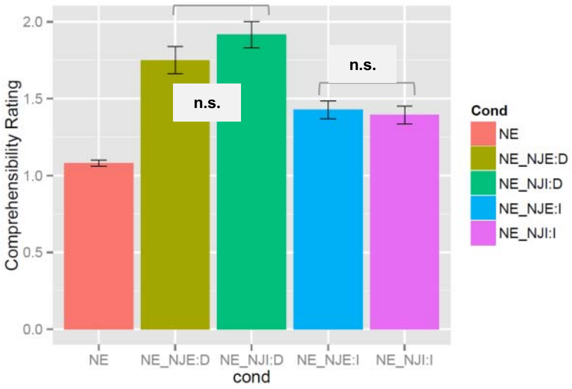


Figure 2: Mean comprehensibility rating scores for each prosody manipulation conditions. (NE: Native English, NE_NJE:D: Native English segments with experienced Japanese durations, NE_NJI:D: Native English segments with inexperienced Japanese durations, NE_NJE:I: Native English segments with experienced Japanese intensity, and NE_NJI:I: Native English segments with inexperienced Japanese intensity). Error bars indicate +/- one standard error.

4. Discussion and conclusions

The aim of this study was to investigate the effect of non-native speech rhythm in the perception of foreign accentedness and comprehensibility. Speech rhythm is typically characterized using duration-based metrics [e.g., 8, 9], but a recent study suggests an important role of amplitude as well [14]. We investigated the influences of both of these properties by transplanting L2 (Japanese speaking English) speech timing and amplitude values onto native (English) segment productions. We also examined how different degrees of L2 English experience (i.e., experienced vs. inexperienced) among the Japanese talkers affected the perceived degree of foreign accent and comprehensibility (i.e., ease of understanding).

The results showed that non-native speech duration affected both accentedness and ease of understanding. That is, the native English listeners perceived L2 duration transplanted sentences as being more foreign-accented and more difficult to understand than untransplanted (control) sentences. The talker's L2 experience had an effect only on the perceived degree of foreign accent, but not on listening effort. This result is consistent with previous research showing that listeners are more sensitive to foreign accent than the effort to understand [3]. We suggest that the less experienced L2 talker's sentences were rated having more foreign accent due to their timing being more variable. In our previous acoustic analyses, we found that L2 vowel duration variability (rhythm) was larger for the inexperienced compare to the experienced L2 talkers [19]. Thus, the increase in perceived accentedness for sentences transplanted with the inexperienced Japanese talker's duration may be due to native listener's sensitivity to changes in durational patterns.

Transplanting non-native speech intensity did not produce the same effect on perception as transplanting duration. For example, although L2 talker's experience made a difference for the perception of duration-transplanted speech, transplanted intensity did not. This difference between timing and intensity may involve the degree to which these factors are apparent in acoustic differences between the two talkers and whether these are perceptually salient. That is, there may be little difference in intensity between the experienced and inexperienced L2 talkers. To our knowledge, there is no study that has examined L2 intensity as a function of L2 experience, and our future analyses will address this. In addition, intensity changes may be less perceptually salient than durational ones. Indeed, the effect of non-native intensity was smaller compared to the duration effect for both accentedness and listener effort, and much smaller compared to segmental contributions found in previous research [23, 24]. It is of course possible that although the behavioural effect is subtle, there could be a large effect in brain-based measures. That is, less efficient neural entrainment may occur with non-native intensity patterns considering recent neurophysiological findings [e.g., 25]. Further research is necessary to address this issue.

Overall, our findings highlight the role of non-native duration patterns on the perceived degree of foreign accent as a function of L2 experience. We are currently conducting experiments examining intensity measures for the English production by the experienced and inexperienced L2 talkers, aiming to understand how L2 speech rhythm is acoustically different from native speech rhythm and how the difference affects perception. We believe that the current results offer an insight into the many ways that L2 rhythm can affect the perception of foreign-accent.

5. References

[1] Munro, M., and Derwing, T., "Foreign accent, comprehensibility, and intelligibility in the speech of second language learners", Language Learning, 45, 73-97, 1995.

[2] Derwing, T. M., and Munro, M.J, "Putting accent in its place: Rethinking obstacles to communication", Language Teaching 42, 276-490, 2009.

[3] Derwing, T. M., and Munro, M. J., "Accent, intelligibility, and comprehensibility: Evidence from four L1s", Studies in Second Language Acquisition, 19, 1-16, 1997.

[4] Miyawaki, K., Jenkins, J. J., Strange, W., Liberman, A. M., Verbrugge, R., and Fujimura, O., "An effect of linguistic experience: The discrimination of [r] and [l] by native speakers of Japanese and English," Perception & Psychophysics, vol. 18, 331–340, 1975.

[5] A. Sheldon, and W. Strange., "The acquisition of /r/ and /l/ by Japanese learners of English: Evidence that speech production can precede speech perception," Applied Psycholinguistics, vol. 3(3), 243-261, 1982.

[6] Iverson, P., and Evans, B. G., "Learning English vowels with different first-language vowel systems: Perception of formant targets, formant movement, and duration" The Journal of the Acoustical Society of America, 122(5), 2842-2854, 2007.

[7] Oh, G. E., Guion-Anderson, S., Aoyama, K., Flege, J. E., Akahane-Yamada, R., and Yamada, T., "A one-year longitudinal study of English and Japanese vowel production by Japanese adults and children in an English-speaking setting", Journal of Phonetics, 39(2), 156-157, 2011.

[8] White, L., and Mattys, S. L., "Calibrating rhythm: First language and second language studies," Journal of Phonetics, vol. 35, (4), pp. 501-522, 2007.

[9] Ordin, M., and Polyanskaya, L., "Development of timing patterns in first and second languages," System, 42, 244-257, 2014.

[10] Aoyama, K., and Guion, S. G., "Prosody in second language acquisition: Acoustic Analyses of duration and F0 range". In Bohn, O.-S., Munro, M.J. (Eds.), Language Experience in Second Language Speech Learning: In Honor of James Emil Flege, Amsterdam, John Benjamins, 282-297, 2007.

[11] Lehiste, I., "Suprasegmentals", Cambridge, MA: MIT Press, 1970.

[12] Abercrombie, D., "Elements of general phonetics" (Vol. 203): Edinburgh University Press Edinburgh, 1967.

[13] Pike, K. L., "The Intonation of American English", 1945.

[14] Leong, V, "Prosodic rhythm in the speech amplitude envelope: Amplitude modulation phase hierarchies (AMPHs) and AMPH models", Cambridge, UK: PhD Thesis, 2012.

[15] 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, York, Timonium, MD, 233–273, 1995.

[16] Flege, J. E., "Interactions between the native and second-language phonetic systems", An Integrated View of Language Development: Papers in Honor of Henning Wode,Trier Wissenschaftlicher Verlag., 217-244, 2002.

[17] Best, C.T. and Tyler, M.D., Nonnative and second-language speech perception: Commonalities and complementarities, in Second language speech learning: The role of language experience in speech perception and production, M.J. Munro and O.-S. Bohn, Eds. 2007, John Benjamins: Amsterdam. 13-34.

[18] Pinet, M., and Iverson, P., "Talker-listener accent interactions in speech-in-noise recognition: Effects of prosodic manipulation as a function of language experience", The Journal of the Acoustical Society of America, 128(3), 1357-1365, 2010.

[19] Kawase, S., Kim, J., and Davis, C., "The influence of second language experience on Japanese-accented English rhythm," Proceeding of Speech Prosody 2016, Boston, USA.

[20] Mochizuki-Sudo, and S. Kiritani., "Production and perception of stress-related durational patterns in Japanese learners of English," Journal of Phonetics, vol. 19(2), 231-248, 1991.

[21] Yoon, K., "Imposing native speakers' prosody on non-native speakers' utterances: The technique of cloning prosody", Journal of the Modern British & American Language & Literature, 25(4), 197-215, 2007.

[22] Hothorn, T., Bretz, F., and Westfall, P., "Simultaneous inference in general parametric models", Biometrical journal, 50(3), 346-363, 2016.

[23] Winters, S, and O'Brien, M. G., "Perceived accentedness and intelligibility: The relative contributions of F0 and duration", Speech Communication, 55 (3), 486-507, 2013.

[24] Rognoni, L., and Busà, M. G., "Testing the Effects of Segmental and Suprasegmental Phonetic Cues in Foreign Accent Rating: An Experiment Using Prosody Transplantation", Proceedings of the International Symposium on the Acquisition of Second Language Speech Concordia Working Papers in Applied Linguistics, 5, 547-560, 2014

[25] Schroeder, C. E., and Lakatos, P., "Low-frequency neuronal oscillations as instruments of sensory selection", Trends in Neurosciences, 32(1), 9-18, 2009.

The Role of Closure Duration in the Perception of Word-Initial Geminates in Kelantan Malay

Mohd Hilmi Hamzah[1], *John Hajek*[2], *Janet Fletcher*[2]

[1]School of Education and Modern Languages, Universiti Utara Malaysia, Malaysia
[2] School of Languages & Linguistics, The University of Melbourne, Australia
hilmihamzah@uum.edu.my, johnth@unimelb.edu.au, janetf@unimelb.edu.au

Abstract

This study examines the extent to which closure duration plays a role in perceptually cueing Kelantan Malay (KM) word-initial geminates. Three word-pairs (/k-kk/, /b-bb/, /l-ll/) embedded in a carrier sentence were chosen for manipulation. In each pair, the closure phase in word-initial consonants was manipulated. Results show that KM listeners shift their perception from singletons to geminates, or vice versa, in lengthened singletons and shortened geminates respectively. The findings support the view that closure duration is a robust acoustico-perceptual cue to word-initial consonant gemination.

Index Terms: closure duration, word-initial geminates, consonant contrast, geminate perception, Kelantan Malay

1. Introduction

It is well established across languages that the singleton/geminate contrast in consonants, either word-initially or word-medially, is distinguished primarily by a difference in closure duration (e.g., [1]). Perceptually, closure duration has also been shown to be a powerful cue to consonant gemination across languages (e.g., Cypriot Greek [2]; Pattani Malay [3]). In the case of word-initial geminates, the perceptual salience of this parameter has been demonstrated in a series of perception experiments in Pattani Malay (henceforth PM), the Malay variety with which KM shares many phonological features. These experiments (e.g., [3]) involved manipulating the closure phase of voiceless stop singletons and geminates embedded in utterance-medial position and determining the perceptual boundary between singleton and geminate consonants, i.e., the duration value that corresponded to a categorical shift in perception.

For instance, in a perception experiment using the manipulated word-pair /paka/ ‘to use’ versus /ppaka/ ‘usable’ [3], it was found that a difference in closure duration is sufficient to cue the PM short versus long distinction; lengthening the word-initial singleton /p/ leads to more responses for the geminate category, while shortening the word-initial geminate /pp/ results in more responses for the singleton category. In both cases, there are perceptual crossovers with complete perceptual shifts to the opposite consonant categories, in line with the acoustic findings for PM [4] and supported by other studies (e.g., [5]) dealing with the perceptual effect of closure duration on consonant gemination.

However, the PM study in [3] also shows that the crossover-points are different between the groups of stimuli; for the stimuli made from the original word with the singleton /p/, the 50% perceptual crossover point is at 120 ms, while for those created from the original word with the geminate /pp/, the perceptual shift begins at 104 ms, i.e., 16 ms earlier than that for the former group of stimuli. The difference between these two crossover points is statistically reliable (p<.001), suggesting that there may be additional acoustic cues at play that influence the placement of the category boundary.

In KM, the robust role of closure duration in characterizing the production of the singleton/geminate contrast has been demonstrated in earlier acoustic studies, e.g., [6]. In the current study, we aim to examine whether and how closure duration also serves as a perceptual cue to KM word-initial geminates among KM native listeners. We are especially interested in examining what contribution, if any, controlled changes in closure duration make to the perception of the word-initial singleton/geminate contrast in KM. In addition, we also aim to identify whether there are ambiguous zones between the stimuli created from original singletons and original geminates that may suggest the presence of other acoustic cues alongside closure duration.

2. Method

2.1. Materials

The word-pairs chosen for manipulation in this study are displayed in Table 1 below. The word-pairs consist of three consonant groups (i.e., voiceless stops /k-kk/, voiced stops /b-bb/ and sonorants /l-ll/). This choice of consonant group is based on Abramson’s methodology in his investigation of word-initial geminates in PM [3]. Since the manipulation of acoustic closure duration for voiceless stop tokens is only possible for those recorded in utterance-medial position, all word-pairs were taken from this environment to ensure consistency.

Table 1. *Sources of stimuli.*

Consonant group	**Singleton**		**Geminate**	
	Word	**Gloss**	**Word**	**Gloss**
Voiceless stops	/kabo/	blurry	/kkabo/	a beetle
Voiced stops	/batʃɔ/	read	/bbatʃɔ/	is reading
Sonorants	/lapu/	lights	/llapu/	on the lights

The procedures for manipulation were achieved by using the manipulation editor in Praat version 5.1.11 [7]. Closure duration was varied along a set of continua in two ways. First, the closure duration values of each singleton were *lengthened* in a series of controlled steps. For the voiceless and voiced stop singletons /k/ and /b/, their duration values were each lengthened from their original values in eleven 10-ms steps, resulting in twelve stimuli each including their original singletons. For the sonorant singleton /l/, its closure duration values were lengthened from the original values in ten 10-ms steps, resulting in eleven stimuli including the original singleton. The duration values of the longest variant for each singleton group were all similar to their original geminate counterparts, i.e., 173 ms for the longest /k/, 180 ms for the longest /b/, and 170 ms for the longest /l/.

Second, the closure duration values of each geminate were *shortened* in a series of controlled steps. For the voiceless and voiced stop geminates /kk/ and /bb/, their duration values were each *shortened* from their original values in eleven 10-ms steps, resulting in twelve stimuli each including their original geminates. For the sonorant geminate /ll/, its closure duration values were shortened from the original values in ten 10-ms steps, resulting in eleven stimuli including the original geminate. The duration values of the shortest variant for each geminate group were all similar to their original singleton counterparts, i.e., 63 ms for the shortest /kk/, 70 ms for the shortest /bb/, and 70 ms for the shortest /ll/. The total number of manipulated stimuli that consisted of all the singleton and geminate variants including their original words was 24 each for the voiceless and voiced stop pairs /k-kk/ and /b-bb/, and 22 for the sonorant pair /l-ll/. Altogether, the manipulation of closure duration for each word in all three pairs yielded 70 stimuli. They were presented three times to the listeners, creating a total of 210 manipulated trials.

2.2. Listeners and Data Collection

The participants for all perception experiments were 30 undergraduate students (15 males, 15 females), all native speakers of KM, at the Universiti Malaysia Kelantan, Kelantan, Malaysia. Their ages ranged between 20 to 25 years (mean age: 21.2). At the time of the experiments, they exhibited no symptoms of hearing disability. All of the listeners were born and raised in Kelantan, Malaysia.

The listeners participated individually in the perception experiment in a quiet room at the Universiti Malaysia Kelantan. They were seated at a desk and were fitted with a stereo headphone in order to listen to the experiment stimuli. All the stimuli were presented through a computer using Praat's Experiment Multiple Forced Choice listening experiment (version 5.1.11). Three experiment files were designed that fitted the three minimal pairs tested in this experiment. These files were played in Praat during the experiment.

The first author gave verbal instructions to the listeners in their native language, i.e., KM. The listeners were first trained with a few stimuli before the experiment took place so that they were comfortable with the experiment design. The listeners then read the instruction on a computer screen, informing them to listen to a sound and choose the word that most closely resembled to what they were going to listen. Since there is no written counterpart of KM, all the words were written in Standard Malay that corresponded to the stimuli in KM.

During the experiment, only one pair was tested at a time. Therefore, the same response categories appeared on the screen for each pair. For each stimulus, there was initial silence duration of 1.5 seconds. The listeners were allowed to replay each stimulus once. The experiment files were run separately but sequentially and controlled by the first author. There was one break after the 36th stimulus for each pair. The experiment lasted for approximately thirty minutes for each participant. All listeners were financially compensated for their participation.

2.3. Data Analysis

The results for each listener were saved in a separate response file. The responses for each listener and each stimulus were processed using Excel (version 14.1.0). Since the stimuli were presented three times in each experiment, the scores for the correct responses ranged from 0 to 3. These scores represented the correct responses for geminates. The total scores for each stimulus were then converted into percentages and plotted into response curves. Geminate responses to each series of stimuli made from original words with singletons or geminates were submitted to one-way ANOVA tests using SPSS (version 20.0.0) to determine their significance levels.

Following [3], the differences observed between the two series of stimuli in the 50% crossover points (i.e., the boundary on a response curve where the listeners' perception switches from singletons to geminates or vice versa) were calculated and compared statistically using ANOVA. Samples paired *t*-tests were also employed to test the level of significance of geminate responses between the two groups of stimuli at a specific step on a duration continuum.

3. Results

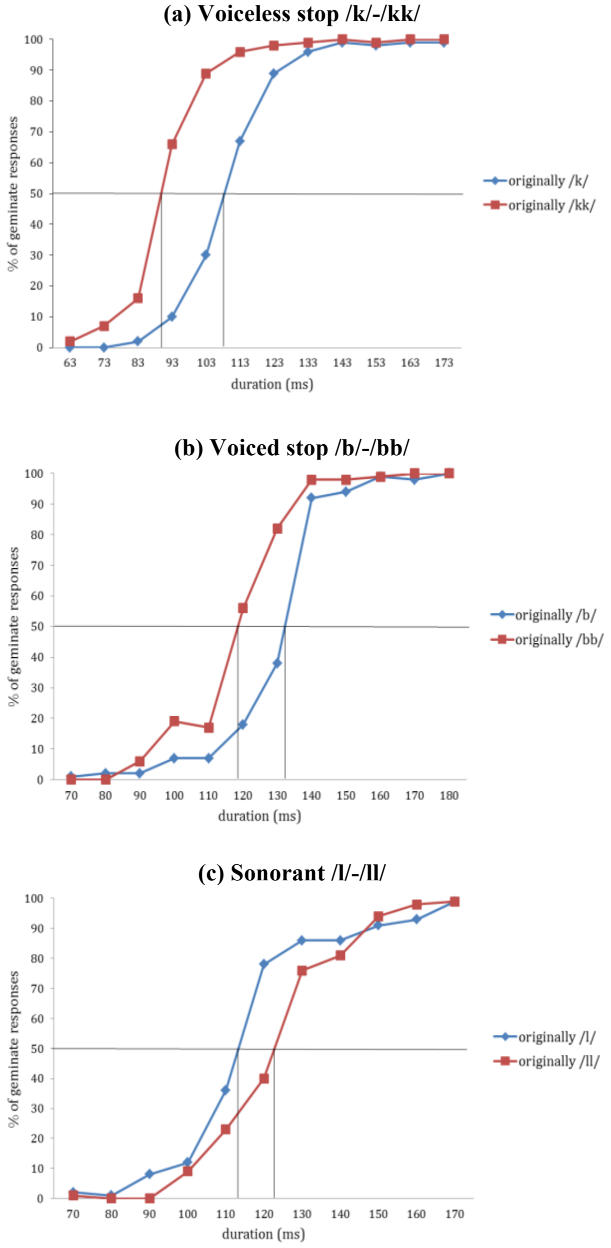
The response curves in Figure 1 illustrate the perception results for manipulated stimuli, showing mean percentages of geminate responses for (a) /kabo/-/kkabo/, (b) /batʃə/-/bbatʃə/ and (c) /lapu/-/llapu/. Detailed measurements underlying these response curves are provided in Table 2. The leftmost durations are the original durations for lengthened singletons, while the rightmost durations are the original durations for shortened geminates. Horizontal lines show the crossover zones between singletons and geminates at 50%, while vertical lines indicate the 50% crossover points between the two series of stimuli.

As shown in the response curves in Figure 1(a-c), there are in general similar patterns across all word-pairs: lengthening of original singletons brings about more geminate responses, while shortening of original geminates causes fewer geminate responses. The response curves are all categorical; in each case, there is a clear perceptual boundary with a single 50% crossover point for each series of stimuli. At the beginning of the duration continua where the original closure durations for words with singletons are located, geminate responses are almost virtually non-existent across word-pairs. That is, almost all the stimuli are identified as singletons. By contrast, at the far right end of the continua where the original closure durations for words with geminates are located, all the stimuli are identified almost 100% of the time as geminates.

For statistical treatment, geminate responses (i.e., the dependent variable) for each series of the stimuli across word-pairs were submitted to one-way ANOVA tests. Results show that the differences are highly significant (*p*<.001) across the board, indicating that closure duration plays a sufficient role in perceptually cueing the consonant length distinction across word-pairs tested in this experiment:

1. original /k/ (F(11,682.7)=209.0, *p*<.001); original /kk/ (F(11,550.0)=206.6, *p*<.001)
2. original /b/ (F(11,688.0)=256.3, *p*<.001); original /bb/ (F(11,687.9)=177.9, *p*<.001)

3. original /l/ (F(10,596.4)=112.1, *p*<.001); original /ll/ (F(10,602.5)=136.4, *p*<.001)


(a) Voiceless stop /k/-/kk/

(b) Voiced stop /b/-/bb/

(c) Sonorant /l/-/ll/

Figure 1: *Mean percentages of geminate responses to the manipulated stimuli.*

Table 2. *Number of tokens and mean percentages of geminate responses to the manipulated stimuli. Significant differences in the responses between the two types of stimuli are highlighted in grey ('***' highly significant; '**' moderately significant; '*' just significant; 'n.s.' not significant).*

(a) Voiceless stop /k/-/kk/

Duration (ms)	n	Originally /k/ (%)	n	Originally /kk/ (%)	Sig.
63	90	0	90	2	0.161 n.s.
73	90	0	90	7	0.056 n.s.
83	90	2	90	16	<0.01 **
93	90	10	90	66	<0.001 ***
103	90	30	90	89	<0.001 ***
113	90	67	90	96	<0.001 ***
123	90	89	90	98	<0.05 *
133	90	96	90	99	0.184 n.s.
143	90	99	90	100	0.326 n.s.
153	90	98	90	99	0.573 n.s.
163	90	99	90	100	0.326 n.s.
173	90	99	90	100	0.326 n.s.

(b) Voiced stop /b/-/bb/

Duration (ms)	n	Originally /b/ (%)	n	Originally /bb/ (%)	Sig.
70	90	1	90	0	0.326 n.s.
80	90	2	90	0	0.161 n.s.
90	90	2	90	6	0.184 n.s.
100	90	7	90	19	<0.05 *
110	90	7	90	17	0.083 n.s.
120	90	18	90	56	<0.001 ***
130	90	38	90	82	<0.001 ***
140	90	92	90	98	0.096 n.s.
150	90	94	90	98	0.264 n.s.
160	90	99	90	99	1.000 n.s.
170	90	98	90	100	0.161 n.s.
180	90	100	90	100	1.000 n.s.

(c) Sonorant /l/-/ll/

Duration (ms)	n	Originally /l/ (%)	n	Originally /ll/ (%)	Sig.
70	90	2	90	1	0.573 n.s.
80	90	1	90	0	0.326 n.s.
90	90	8	90	0	0.070 n.s.
100	90	12	90	9	0.448 n.s.
110	90	36	90	23	0.102 n.s.
120	90	78	90	40	<0.001 ***
130	90	86	90	76	0.054 n.s.
140	90	86	90	81	0.214 n.s.
150	90	91	90	94	0.522 n.s.
160	90	93	90	98	0.255 n.s.
170	90	99	90	99	1.000 n.s.

However, as shown in Figure 1 and Table 2, the identification curves for the manipulated stimuli with lengthened singletons and shortened geminates are not entirely identical. It can be observed across word-pairs that there are ambiguous zones between the two response curves, as indicated by two vertical lines in each figure. In the case of the voiceless stop pair /k/-/kk/ and the voiced stop pair /b/-/bb/, the perceptual crossovers at 50% from the singleton to the geminate category are earlier for the stimuli made from original geminates than for those made from original singletons: the duration differences between the two crossover points are 18 ms (for the voiceless stop pair) and 15 ms (for the voiced stop pair). Interestingly, for the sonorant pair /l/-/ll/, the crossover is unexpectedly earlier for the stimulus with the original singleton /l/ rather than for that with the original geminate /ll/: the duration difference is 10 ms between the two crossover points. Note that the ambiguous zone is largest for the voiceless stop pair (18 ms), followed by the voiced stop pair (15 ms) and then the sonorant pair (10 ms). ANOVA tests indicate that the differences between crossover points are significant across all word-pairs: /k/-/kk/ (F(1,540)=18.588, *p*<.001); /b/-/bb/ (F(1,540)=15.422, *p*<.001); /l/-/ll/ (F(1,540)=10.558, *p*<.01).

Additionally, geminate responses to different series of stimuli appear to diverge significantly at several duration points on the duration continua, as highlighted in grey in Table 2(a-c). For the voiceless stop pair /k/-/kk/, significant differences occur at five consecutive duration points: 83 ms (*p*<.01), 93 ms (*p*<.001), 103 ms (*p*<.001), 113 ms (*p*<.001) and 123 ms (*p*<.05). As for the voiced stop pair /b/-/bb/, the differences are significant at three points: 100 ms (*p*<.05), 120 ms (*p*<.001) and 130 ms (*p*<.001). Note the confusion at the duration point of 110 ms, as indicated by the zigzag line in Figure 1(b), in which listeners' geminate responses to the stimuli made from the original /bb/ drop slightly by 2%, although the differences are not statistically significant (*p*=0.083). Finally, in the case of the sonorant pair /l/-/ll/, the differences are only significant at one duration point, i.e., 120 ms (*p*<.001).

4. Discussion and Conclusions

In this study, we have looked at the effect of closure duration on the perception of the word-initial singleton/geminate contrast in KM. In doing so, we have examined carefully modified stimuli with incremented/decremented closure durations placed in utterance-medial contexts. The results show that closure duration is a highly significant acoustico-perceptual cue to KM word-initial geminate consonants. It appears that KM listeners respond reliably to opposite categories when the duration continuum shifts in a series of controlled steps, i.e., lengthened singletons bring about more geminate responses while shortened geminates lead to more singleton responses. The near-complete perceptual crossovers between singletons and geminates demonstrated across the manipulated stimuli made across word-pairs can be interpreted as strongly supporting the acoustic findings on closure duration for KM presented in [6], in which the durational contrast between singletons and geminates is evident in the consonant closures across all contexts in all phoneme categories in KM.

It appears that listeners are sensitive to closure duration information for all tested word-pairs so much so that, at the extreme ends of the duration continua, closure duration overrides any other possible cues, such as unmanipulated amplitude or F0, giving almost complete judgments of opposite categories, i.e., either singletons or geminates, depending on whether the manipulated stimuli are made from original singletons or geminates. This observation is comparable to Abramson's experiments with synthesized stimuli in PM [3]. More importantly, this observation is broadly in agreement with the universal claim in the literature on the robust role of closure duration in defining consonant gemination (e.g., [8]).

Our results have also shown that, despite the powerful effect of closure duration, there are displacements at the category boundary between the two series of stimuli. In KM, there is an 18-ms difference between the crossover points for the stimuli created from the original /k/ and /kk/, which is almost identical to that reported in PM [3] between the stimuli made from the original /p/ and /pp/ (the crossover-point difference is 16 ms, *p*<.001). Collectively, the displacements at crossover points shown in KM and PM are larger than in languages with word-medial geminates, such as in Bengali [5] in which the difference between crossover points between two series of stimuli (i.e., original singletons and original geminates) is reported to be 10 ms for /t/-/tt/. All other things being equal, this cross-linguistic difference is possibly due to the fact that consonant gemination in KM and PM only occurs in word-initial position which, in the case of voiceless stops, can be reliably accompanied by additional acoustic cues (e.g., VOT differences) even in utterance-medial position where closure duration cue is also present, as shown in [9,10,11] for KM.

Another critical point regarding the boundary displacement is that, in the case of voiceless stop stimuli, an earlier crossover in KM is exhibited in the stimuli made from the original word with a geminate (i.e., /kk/), which is again consistent with the data for PM [3]; the 50% crossover point is earlier for the stimuli synthesized from the original geminate /pp/ in PM. As for the case of the sonorant pair /l/-/ll/ in KM, the earlier crossover for the stimulus with the original singleton /l/ could be due to the presence of other acoustic correlates (e.g., increased amplitude in the following vowel) associated with its geminate counterpart, as shown in [10] for KM. All in all, the significant differences observed between the two crossover points across all word-pairs suggest that other durational or non-durational acoustic correlates (e.g., F0 differences in the following vowel) may also potentially cue the word-initial singleton/geminate contrast in KM. At this stage, it seems that KM listeners may be attuned to a range of acoustic cues associated with the word-initial consonant contrast in KM in addition to closure duration. Further perception experiments are needed, however, in order to verify this claim.

5. Acknowledgements

The authors would like to thank Universiti Utara Malaysia for its financial support. This work was supported by the university grant S/O code no. 13392.

6. References

[1] Lahiri, A., and Hankamer, J., "The timing of geminate consonants", JPhon, 16, 327-338, 1988.

[2] Muller, J.S., "The production and perception of word-initial geminates in Cypriot Greek", Proc. 15th ICPhS, 1867-1870, 2003.

[3] Abramson, A.S., "The perception of word-initial consonant length: Pattani Malay", JIPA, 16, 8-16, 1986.

[4] Abramson, A.S., "Word-initial consonant length in Pattani Malay", Proc. 11th ICPhS, 68-70, 1987.

[5] Hankamer, J., Lahiri, A., and Koreman, J., "Perception of consonant length: Voiceless stops in Turkish and Bengali", JPhon, 17, 283-298, 1989.

[6] Hamzah, M. H., Fletcher, J., and Hajek, J. "Closure duration as an acoustic correlate of the word-initial singleton/geminate consonant contrast in Kelantan Malay", JPhon, 58, 135–151, 2016.

[7] Boersma, P., "Praat, a system for doing phonetics by computer", Glot International, 5, 341-345, 2001.

[8] Ridouane, R., "Geminates at the junction of phonetics and phonology", in C. Fougeron, B. Kuhnert, M. D'Imperio, and N. Vallée (Eds.), Laboratory Phonology 10, 61–90, Mouton, 2010.

[9] Hamzah, M. H., Hajek, J., and Fletcher, J., "A taste of prosody: Possible effects of the word-initial singleton-geminate contrast on post-consonantal vowel duration in Kelantan Malay", Proc. 6th Speech Prosody, 490-493, 2012.

[10] Hamzah, M. H., Fletcher, J., and Hajek, J., "Amplitude and F0 as acoustic correlates of Kelantan Malay word-initial geminates", Proc. 15th Australasian International Conference on SST, 63-66, 2014.

[11] Hamzah, M. H., Fletcher, J., and Hajek, J., "Word-initial voiceless stop geminates in Kelantan Malay: Acoustic evidence from amplitude/F0 ratios", Proc. 18th ICPhS, 2015.

Does a Vowel by Any Other Accent Sound the Same … to Toddler Ears?

Catherine T. Best[1,2], *Christine Kitamura*[1], *Sophie Gates*[1] *and Angela Carpenter*[3]

[1]Western Sydney University, Australia
[2]Haskins Laboratories, USA
[3]Wellesley College, USA

{c.best, c.kitamura, s.gates}@westernsydney.edu.au, acarpent@wellesley.edu

Abstract

Research on spoken word recognition in young children has emphasized detection of minimal phonetic contrasts, and offers conflicting evidence about the role of consonants versus vowels. The complementary ability to recognize words across natural phonetic variation, *phonological constancy*, is equally important to language development. Prior studies of phonological constancy found that 15- and 19-month-olds recognize familiar toddler words in an unfamiliar regional accent containing Category Goodness vowel and/or consonant differences from their native accent, or Category Shifting consonant differences. In the present study, Category Shifting vowel differences disrupted word recognition at both ages, supporting different roles for vowels than consonants.

Index Terms: early word recognition, regional accent variation, perceptual assimilation, vowels versus consonants

1. Introduction

Over four decades of research on infants' perceptual attunement to their native language have provided many insights into experiential effects on speech perception over the course of the first year. Infants under 8 months discriminate many non-native consonant contrasts, but show a dramatic decline in discriminating many, though not all, of these contrasts by 9-10 months. The decline in discrimination appears earlier for vowel contrasts, by around 5-6 months [e.g., 1-4]. Conversely, for some native consonant contrasts, discrimination improves over the first year or so [5, 6]. Importantly, however, some non-native contrasts continue to be discriminated well even into adulthood, suggesting that changes in perception of non-native speech distinctions cannot be attributed solely to lack of early exposure to their specific surface-level acoustic-phonetic properties. Rather, as posited by the Perceptual Assimilation Model (PAM) [7, 8], developmental shifts in perception of both native and non-native contrasts must reflect the emerging ability to map varying phonetic details to more abstract native phonological structures, such as spoken words and their component phonological elements. Thus, when a listener perceives the members of a non-native phonetic distinction as equivalent exemplars of the same single native phoneme (Single Category assimilation: SC), discrimination is poor because they perceive the phonetic difference as not phonologically or lexically contrastive in their language. But if they perceive a non-native distinction as corresponding to a native phonological contrast (Two Category assimilation: TC), they continue to discriminate it well. Moreover, they remain sensitive to some within-category phonetic variation if they detect it as a difference in goodness of fit to a single native phoneme, which they discriminate moderately well though not as well as a TC contrast (Category Goodness: CG) [9, 10].

An important theoretical issue not considered in the cross-language perception literature is the role that perceptual assimilation may play in children's growing ability to recognize words (and their component consonants and vowels) across various types of natural phonetic variations that they encounter *within* their native language. Recognizing that critical phonetic differences can convey phonological distinctions between native words, e.g., that /**p**iz/ (***p**eas*) is not the same word as /**k**iz/ (***k**eys*), is obviously crucial to language development. However, of equal or greater importance is that the child also needs to develop the ability to recognize words across the lexically-irrelevant phonetic variations presented by different speakers and regional accents of their language, i.e., to recognize the *phonological constancy* of words [1, 11-13]. Phonological constancy requires perceptual assimilating those types of variations to their common, underlying phonological forms.

Less is known about the emergence of phonological constancy, but research is growing on the topic. The first such study examined 15- and 19-month-olds' recognition of familiar words (i.e., known to toddlers), as indexed by a preference for listening to sets of toddler words over sets of unfamiliar words (i.e., low-frequency adult words), in their own native regional English accent versus in a phonetically-differing regional English accent they had not previously been exposed to. Whereas both age groups showed a listening preference for the familiar toddler words, only the 19-month-olds showed this preference when tested with the unfamiliar accent, indicating that they but not the younger children had achieved phonological constancy [11]. A follow-up eye-tracking study assessed whether the same age groups could identify the familiar toddler words in a visual preference task, by looking more at the named object than the unnamed one (distractor) in pairs of photographs. Both ages showed a reliable looking preference for words in their native accent, but again only the 19-month-olds did so when the words were spoken in the unfamiliar accent [13]. Two subsequent studies found a relationship between children's expressive vocabulary size and their listening preference for familiar toddler words under task conditions with high stimulus variability (more speakers, words and tokens than in [11]). 15- and 17-month-olds with small vocabularies (≤ 25 words) failed to show a reliable preference for toddler words in either accent, whereas 17-month olds with larger vocabularies (≥ 50 words) showed a familiar toddler word preference in their native accent and 19-month-olds with even larger vocabularies (≥ 100 words) preferred familiar toddler words in both their native and the unfamiliar accent [14, 15]. Thus, vocabulary development is linked to recognizing the phonological constancy of words across moderately high phonetic variation, first in the native accent, and later for the greater variations of an accent not previously experienced.

PAM has since been extended to predict assimilation patterns for regional accent variation in the phonetic realizations of vowels versus consonants within the native language, as perceived by both adults [16-18] and toddlers [19-22]. For many native phonemes regional accent variations are assimilated as Native-Like (NL), i.e., the phonetic deviation from the native accent realization is small enough as to be perceptually insignificant. For others, however, the phonetic deviation is more noticeable as a Category Goodness (CG) difference within the matching native phoneme, i.e., perceived as the correct consonant or vowel but also heard as being pronounced differently than in the native accent. In striking distinction, though, some accent differences may transgress native-accent category boundaries, and thus be perceived as a different consonant or vowel than the other-accent speaker intended, i.e., the difference is Category Shifting (CS). CG versus CS cross-accent assimilations are expected to have notably different influences on recognition of spoken words in the other, non-native accent: correctly though perhaps more slowly identified for CG differences, but incorrectly identified, even by adults, for CS differences. Moreover, the developmental differences for discrimination of non-native vowel versus consonant contrasts summarized earlier suggest that these effects may differ for vowel versus consonant deviations from the native accent. Importantly, the relative role of consonants versus vowels in early word learning is under debate. For example, 11-16-month-olds have been reported to rely more on consonants than vowels [23], or to show symmetrical sensitivities to vowels and consonants [24], in recognition of known words, and have been claimed to rely more on vowels [25], or more on consonants [26], in learning new words.

Given that debate and the previous mixed findings, we ran a series of cross-accent word listening preference studies examined the impact of CG versus CS differences from the children's native accent (Australian English: AusE) in an unfamiliar regional accent's vowel pronunciations (JaME: Jamaican Mesolect English) or consonant pronunciations (London "Cockney" English: CknE). In those studies, for the first time, not only 19- but also 15-month-olds showed a familiar toddler word preference in both AusE and the unfamiliar accent, when the word sets showed only CG differences in either the vowels (JaME) or the consonants (CknE). Conversely, when there were CS differences in both consonants and vowels (JaME), neither age recognized the familiar toddler words in the non-AusE accent. Strikingly, however, when only the consonants showed CS differences (CknE), both age groups did generalize the familiar toddler word preference to the unfamiliar accent [19-22]. Together, those findings indicate, firstly, that even by 15 months, toddlers assimilate CG vowel variations, and both CG and CS consonant variations, in unfamiliar accents to their (native-accented) representations of known words. Secondly, in light of that pattern, the failure of both age groups to recognize words across CS differences in both consonants and vowels implies that their difficulty was due specifically to the CS *vowel* differences, not to the CS consonant differences. But this inference was not directly assessed. Moreover, the two ages may well differ in how they respond to CS vowel differences alone. Therefore, the current study tested whether restricting the pronunciation differences to CS vowel deviations (JaME) from the native accent (AusE) would disrupt 15- and/or 19-month-olds' recognition of familiar toddler words.

2. Method

The virtually identical findings for toddlers' visual identification of words via eye-tracking [13] and for their listening preferences in other studies of cross-accent word recognition [11, 14, 15, 19-22] imply that both tasks index lexical recognition. Therefore, we used the listening preference task for comparison to the other CG-CS vowel and consonant studies, using CS vowel-differing words in AusE versus JaME. We again compared children at 13-15 months ("15 month-olds"), i.e., the early word-learning period (< 25 word expressive vocabulary), and 18-20 months ("19-month-olds"), who have typically reached a 50+ word vocabulary ('vocabulary spurt').

2.1. Participants

Two sets of participants successfully completed the listening preference test in both target accents. The younger group ('15 months') had 32 children (M_{age} = 14.05 mo, range = 13.28 – 15.22 mo; 17 females), as did the older group ('19 months': M_{age} = 19.19 mo, range = 18.51 – 20.22 mo; 17 females). All were full-term at birth, healthy on the test day, lacked familial speech/language disorders, and received little to no exposure to other languages or non-AusE accents including JaME.

30 additional children at the younger age, and 36 at the older age, were tested but excluded due to fussing/crying (*n* = 52), falling asleep (*n* = 2), climbing out of the parent's lap during testing (*n* = 1), parental interference (*n* = 6) or withdrawal from testing (*n* = 1), or technical problems (*n* = 4). This rejection rate (~50%) is typical of 1-2 year olds across a range of tasks, commensurate with their general behavioral tendencies.

2.2. Stimulus materials

The target items were multiple audio tokens of words produced by several speakers of each comparison accent, for each of the two main word types. The listening preference task also uses static visual stimulus displays, such that participants could control audio stimulus presentations by fixating their gaze on the display.

2.2.1. *Visual fixation displays*

The visual fixation displays were colored checkerboards against a white background, with a central circular "swirl" to attract infants' attention to the center of the screen. Two different checkerboard colors (magenta, blue) were used for each child's AusE vs. JaME preference tests, with the color assignment to each test counter-balanced across participants.

2.2.2. *Spoken word audio stimuli*

The listening preference task presents separate, alternating trials containing sets of words known to toddlers vs. sets of low frequency adult words they are highly unlikely to have ever heard (see Table 1). Each word set (Familiar [toddler]; Unfamiliar [adult]) comprised 8 monosyllabic and 8 bisyllabic target words, carefully selected such that: a) Familiar words occur in ≥50% of AusE expressive vocabularies at 13-15 months [27] and/or appear often in toddler picture-books, whereas Unfamiliar words occur ≤ 2/million in standard and Australian English lexical databases [28, 29]; b) each word contained one stressed vowel that differed in a Category-Shifting way between AusE and JaME, i.e., adult AusE listeners hear the JaME vowel as a categorically different one than the speaker intended. All other phonemes in each target word displayed only NL or CG differences between AusE-JaME.

We recorded three female native speakers each of AusE (from western Sydney, Australia) and of JaME (from St Catherine and St. James parishes, Jamaica). We paired speakers across the two accents to assure similar voice quality, F0 mean and range across the AusE and JaME stimulus sets (one JaME speaker's F0 was raised ~10 Hz via Praat resynthesis to better match her AusE pair). Each speaker produced multiple tokens of each target word, using a Shure SM10A headset microphone connected to a Sony PCMM1 portable DAT recorder (44.1 kHz sampling rate). The printed targets were shown in quasi-random order on a laptop screen. For the task, two tokens per target word were selected from each speaker, resulting in a total of 192 tokens for each regional accent (16 words x 2 word sets x 2 repetitions x 3 speakers). Thus, a child was highly unlikely to hear any token more than once in the tests.

Table 1. *Target words for each word set in each accent.*

Single syllable		**Two syllable**	
Familiar (toddler)	**Unfamiliar (adult)**	**Familiar (toddler)**	**Unfamiliar (adult)**
ball	shawl	baby	frailty
bear	mare	doggy	blobby
boat	dose	flower	doubter
bus	shun	grandma	vanguard
door	gore	lolly	fauna
duck	muck	mummy	putty
nose	foes	paper	taper
socks	knocks	water	spotty

2.3. Procedure

We used a word-type listening preference task [30, 31] in which a series of alternating trials plays out a set of familiar (toddler) words versus a set of unfamiliar (adult) words, for as long as the child fixates on the checkerboard display directly in front of them. When the child looks away for ≥ 2 seconds, the trial ends and the checkerboard flashes on/off until the child looks back, at which time the display stabilizes again and the next trial begins. A trained observer monitors child fixations via a hidden low-light video camera below the checkerboard. The child sits in the parent's lap (who is instructed not to point or interfere, and listens to vocal music over Sennheiser HD650 circumaural headphones to mask test stimuli). A significant listening preference (higher fixation times) for the familiar word trials relative to the unfamiliar word trials is taken to index recognition of the familiar toddler words.

Each participant completed two preference tests of eight trials per test (four trials per word set, in alternating trials). In one test all words were spoken in the native AusE accent; in the other they were all spoken in the unfamiliar JaME accent (test order counterbalanced across participants), as in previous cross-accent word recognition studies [11, 14, 15, 19-21].

3. Results

Total fixation times were summed across familiar word trials, and separately across unfamiliar word trials. These data were submitted to a 3-way Analysis of Variance (ANOVA) on the factors age group (15 vs 19 months) x word set (familiar vs unfamiliar) x accent (AusE/native vs. JaME). (An initial 4-way ANOVA found no significant effects of accent order).

A significant main effect of word type, $F_{(1,62)} = 14.030$, p < 0.0001, indicates an overall listening preference for familiar toddler words over unfamiliar adult words. The main effect for accent, $F_{(1,62)} = 4.927$, p = 0.03, revealed significantly greater overall fixation during the AusE test than the JaME test. However, these were qualified by a significant word set x accent interaction, $F_{(1,62)} = 9.742$, p = 0.003, showing that the familiar word preference was reliable only for AusE, not for JaME (see Figure 1). This held true for both ages: the main effect of age was not significant, nor did it interact with word set or accent.

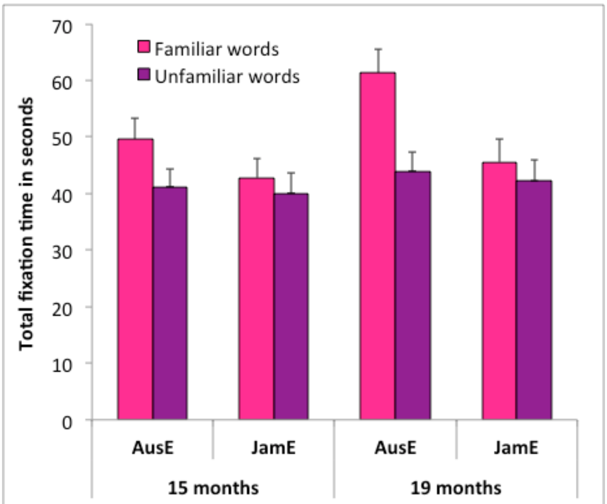


Figure 1: *Total listening times to familiar toddler vs. unfamiliar adult word sets spoken in the native accent (AusE) vs. the unfamiliar accent (JaME), for younger (15 months) and older toddlers (19 months). Error bars are standard errors of the means (s.e.m.).*

4. Discussion

The current findings are compatible with the picture of results from prior investigations of toddlers' ability to recognize familiar toddler words spoken in an unfamiliar accent when the differences from the native accent (AusE) were restricted to either a CG or a CS difference in vowels and/or consonants [19-22]. Here, we found that word recognition was disrupted for the unfamiliar JaME accent in both the 15- and 19-month groups when the JaME pronunciation showed a single CS vowel difference from their native AusE accent. By comparison, the previous studies had found word recognition to remain intact for an unfamiliar accent that differs from native AusE in CG vowel differences alone or in either CG *or* CS consonant differences, but that CS differences in vowels *and* consonants disrupts cross-accent recognition of familiar toddler words. The latter finding, then, appears to have been due solely to the CS vowel differences, and not to the CS consonant differences. Analogously, AusE-speaking adults show marked differences in how they perceptually assimilate vowel versus consonant differences of other English regional accents relative to AusE [16-18]. These cross-accent findings thus extend, and offer some challenges to, previous reports on differential roles for vowels and consonant in young children's learning and recognition of spoken words [23, 25, 26, *cf* 24].

Note, however, that this research has thus far only examined CG and CS vowel versus consonant differences among regional accents of English. In English, vowel pronunciation differences provide the primary source of regional accent variation; consonant differences are much more restricted in English. There are languages in which regional variation instead involves more consonant than vowel variation, for example as a result of variable consonantal lenition processes, such as Spanish. Examining these effects across accents of such languages will be important to teasing out the basis for differential tolerance of the two types of pronunciation variations.

5. Conclusions

We speculate that perceptual assimilation operates not only for non-native speech perception but also within the listeners'

Table of Contents for this manuscript

native language. It reflects the necessity to handle phonological abstraction across natural, systematic phonetic variation within the language, and is evident in the early development of spoken word recognition once the child has achieved phonological constancy. Furthermore, this developmental achievement appears to proceed differently for accommodating Category Goodness (CG) differences and Category Shifting (CS) differences between the native accent and other unfamiliar regional accents, and in particular for CS vowel versus consonant differences. But these effects have only been investigated in English, in which regional accents differ more in vowel than consonant pronunciations. Research on languages with more consonantal than vowel regional variation are needed to evaluate the extent to which, and conditions under which, vowel versus consonant effects in spoken word recognition reflect universal versus language-specific principles.

6. Acknowledgements

Research supported by Australian Research Council grants DP0772441 and DP130104237.

7. References

[1] Best, C. T. “Devil or angel in the details? Complementary principles of phonetic variation provide the key to phonological structure”, in J. Romero and M. Riera [Eds] Sounds, representations and methodologies: Essays on the phonetics-phonology interface, 3-31. John Benjamins, 2015.

[2] Best, C. T. “Speech perception in infants: Propagating the effects of language experience”, in E. M. Fernandez & H. S. Cairns [Eds], Handbook of psycholinguistics. Wiley, in press.

[3] Best , C. T. “Learning to perceive the sound pattern of English”, in C. Rovee-Collier and L. Lipsitt [Eds], Advances in infancy research, 217-304, Ablex, 1994.

[4] Maurer, D. and Werker, J. F. “Perceptual narrowing during infancy: A comparison of language and faces”, Developmental Psychobiology, 56(2):154-178, 2014.

[5] Kuhl, P. K., Stevens, E., Hayashi, A., Deguchi, T., Kiritani, S. and Iverson, P. “Infants show a facilitation effect for native language phonetic perception between 6 and 12 months”, Developmental Science, 9:F13–F21, 2006.

[6] Polka, L., Colantonio, C. and Sundara, M. “A cross-language comparison of /d/-/ð/ perception: Evidence for a new developmental pattern”, J. Acou. Soc. Am., 109:2190-2201, 2001.

[7] Best, C. T. “A direct realist view of cross-language speech perception”, in W. Strange [Ed], Speech perception and linguistic experience, 171–206, York Press, 1995.

[8] Best, C. T. and Tyler, M. D. Nonnative and second-language speech perception: Commonalities and complementarities”, in M. Munro and O-S. Bohn [Eds], Second language speech learning, 13-34, John Benjamins, 2007.

[9] Best, C. T., McRoberts, G. W. and Sithole, N. M. “Examination of perceptual reorganization for nonnative speech contrasts: Zulu click discrimination by English-speaking adults and infants”, J. Exp. Psych.: Human Perc. Perf., 14:45-60, 1988.

[10] Best, C. T., McRoberts, G. W. and Goodell, E. “American listeners' perception of nonnative consonant contrasts varying in perceptual assimilation to English phonology”, J. Acou. Soc. Am., 1097:775-794, 2001.

[11] Best, C. T., Tyler, M. D., Gooding, T. N., Orlando, C. B. and Quann, C.A. **“**Development of phonological constancy: Toddlers’ perception of native- and Jamaican-accented words”, Psych. Science, 20:539-542, 2009.

[12] Watson, T. L., Robbins, R. A. and Best, C. T. “Development of face and speech perception: An integrated approach”, Developmental Psychobiology, 56:1454-1481, 2014.

[13] Mulak, K. and Best, C. T. “Development of word recognition across speakers and accents”, in L. Gogate and G. Hollich [Eds], Theoretical and computational models of word learning, 242–269, IGI Global-Robotics, 2013.

[14] Best, C. T., Tyler, M. D. Kitamura, C., Notley, A. and Bundgaard-Nielsen, R. “Phonetic specificity of early words? Australian toddlers’ perception of Australian versus Jamaican English pronunciations”, Presented at International Conference on Infant Studies, Vancouver, March 2008.

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

[16] Best, C. T., Shaw, J., Mulak, K., Docherty, G., Evans, B., Foulkes, P. Hay, J., Al-Tamimi, J., Mair, K., and Wood, S. Perceiving and adapting to regional accent differences among vowel subsystems. *Proceedings of the 18th International Congress of Phonetic Sciences*, Glasgow, 2015.

[17] Best, C. T., Shaw, J., Mulak, K., Docherty, G., Evans, B., Foulkes, P. Hay, J., Al-Tamimi, J., Mair, K., and Wood, S. “From Newcastle MOUTH to Aussie ears: Australians’ perceptual assimilation and adaptation for Newcastle UK vowels”, *Proceedings of Interspeech*. Dresden, 2015.

[18] Shaw, J., Best, C. T., Mulak, K., Docherty, G., Evans, B., Foulkes, P. Hay, J., Al-Tamimi, J., Mair, K., Peek, M. and Wood, S. “Effects of short-term exposure to unfamiliar regional accents: Australians’ categorization of London and Yorkshire English consonants”, Presented at Australasian Speech Science and Technology, Christchurch, December, 2014.

[19] Best, C. T., Kitamura, C., Pal, M. and Dwyer, A. “Young toddlers recognize non-native Jamaican-accented words differing only in ‘category-goodness’ from native-accent vowel pronunciations”, presented at International Conference on Infant Studies, Minneapolis, June 2012.

[20] Best, C. T., Gates, S., Kitamura, C., Docherty, G., Pinet, M. and Evans, B. G. “Young Australian toddlers recognize Cockney-accented words involving only ‘category-goodness’ differences in consonant pronunciations from Australian English”, Presented at International Conference on Infant Studies, Berlin, July 2014.

[21] Best, C. T. and Kitamura, C. The role of perceptual assimilation in early development of recognition of words spoken in native vs unfamiliar regional accents”, Presented at International Conference on Infant Studies, Berlin, July 2014).

[22] Best, C. T., Gates, S., Kitamura, C., Docherty, G. and Evans, B. “Category-shifting consonant differences between English accents do not interfere with familiar word recognition at 14 or 19 months”, Presented at International Conference in Infant Studies. New Orleans, May 2016.

[23] Hochmann, J-R., Benavides-Varela, S., Nespor, M. and Mehler, J. “Consonants and vowels: Different roles in early language acquisition”, Dev. Sci., 14(6):1445-1458, 2011.

[24] Mani, N., and Plunkett, K. “Phonological specificity of vowels and consonants in early lexical representations”, J. Mem. Lang., 57(2):252-272, 2007.

[25] Pons, F., and Toro, J. M. “Structural generalizations over consonants and vowels in 11-month-old infants”, Cognition, 116(3):361-367, 2010.

[26] Havy, M., and Nazzi, T. “Better processing of consonantal over vocalic information in word learning at 16 months of age”, Infancy, 14(4):439-456, 2009).

[27] Kalashnikova, M., Schwarz, I. C. and Burnham, D. “OZI: Australian English Communicative Development Inventory”, First Language, 0142723716648846, 2016.

[28] Baayen, R. H., Piepenbrock, R., & Van Rijn, H. “The CELEX lexical data base, release 2 [CD-ROM]”, Linguistic Data Consortium, 1995. Available online: http://www.ldc.upenn.edu/

[29] Dennis, S. “The Sydney Morning Herald Word Database”, Noetica: Open Forum, 1995. http://psy.uq.edu.au/cogpsych/noetica/.

[30] Hallé, P. A., and de Boysson-Bardies, B. “Emergence of an early receptive lexicon: Infants' recognition of words”, Infant Behavior and Development, 17(2):119-129, 1994.

[31] Hallé, P. A., & de Boysson-Bardies, B. “The format of representation of recognized words in infants' early receptive lexicon”, Infant Behavior and Development, 19(4):463-481, 1996.

Measuring sensitivity to phonological detail in monolingual and bilingual infants using pupillometry

Katalin Tamási[a], Thilanga D. Wewalaarachchi[b], Barbara Höhle[a] and Leher Singh[b]

[a]International Doctorate for Experimental Approaches to Language and Brain
[b]National University of Singapore, Department of Psychology
{tamasi,hoehle}@uni-potsdam.de, thilanga.d.w@gmail.com, leher.singh.nus@gmail.com

Abstract

This study shows for the first time that mispronunciation detection in 24-month-old mono- and bilingual children can be assessed using pupil data from a preferential looking study. Mispronounced words (specifically, consonant, vowel and tone changes) resulted in larger degrees of pupil dilation for bilingual children than correctly produced words, whereas monolingual children's pupillary responses provided no evidence of sensitivity to mispronunciations. Between-group comparisons revealed that pupil dilation of bilingual children in response to correct labels was lower than monolingual children. Overall performance reversed with consonant-changed labels, but was comparable with vowel- and tone-changed labels. Taking degree of pupil dilation as a proxy of cognitive effort, we argue that in comparison to monolinguals, bilingual children seem to require fewer resources for processing correct labels, while if anything, more resources for processing mispronounced labels in order to activate the corresponding lexical entry. This finding – converging on past research showing enhanced bilingual sensitivity to detect mismatch and mispronunciations [1] – further supports the notion that certain aspects of bilingual early words are represented in greater detail than those of monolingual words.

Index Terms: bilingualism; mispronunciation detection; lexical development; eye tracking; pupillometry

1. Introduction

One of the first steps in language acquisition is to identify the building blocks of words in the target language by segmenting the auditory input and categorizing the resulting sounds into discrete groups of phonemes. This process of phoneme categorization is intricately intertwined with word learning and word recognition. The ability to detect a small yet contrastive change is crucial to building up a rich lexicon that contains minimally different words (e.g., *tap* and *cap*). Previous research tested children's lexical knowledge by presenting them with correct and incorrect forms of words (e.g., *ball* and *dall*), showing that between 11 and 24 months, children learn to differentiate between the two (preference for correct labels in head-turn preference paradigms: [2]; noticing phonemic change in habituation paradigms: [3], longer target looks with correct label in preferential looking paradigms: [4]; difference in event-related brain potential signature: [5]). These findings suggest that early lexical representations are specific, i.e., containing sufficient information to allow the learner to match with the input.

The majority of mispronunciation detection studies to date have relied on looking behavior. Eye tracking studies yield a valuable body of data, only a fraction of which are routinely considered. This study extends the methodology conventionally used in eye tracking studies by complementing the preferential looking paradigm with a measure automatically collected using such a paradigm: pupil dilation. Pupil dilation in children has been linked to cognitive effort, surprise, and novelty [6], making it an appealing tool for infant research. Recently, pupillometry was demonstrated to be sensitive to acoustic (dis-)similarity: [7], semantic violation: [8], and mispronunciations: [9, 10, 11].

Using single-picture pupillometry paradigms – presenting a single visual stimulus per trial –, 30-month-old children have been shown to give a differential pupillary response to correctly pronounced labels and their mispronunciations: the general finding being that mispronounced labels were associated with larger degrees of pupil dilation than correct labels [9, 10]. This asymmetry was interpreted such that more cognitive effort was needed to establish the link between the mispronounced label and the picture (in order to reconstruct the correct form and map onto the corresponding lexical representation) than doing so with the correct label. Such finding is consistent with earlier studies that demonstrated the specificity of lexical representations [2, 3, 4, 5].

Most recently, 30 month-old children's pupil dilation data collected from a preferential looking paradigm have been found to exhibit a linear trend in response to the degree of mispronunciation [11]. The methodological import of the study is showing that pupillary response data can act as an additional measure to looking behavior in preferential looking paradigms as even in the presence of the distractor picture, the degree of pupil dilation remains associated with the cognitive effort of reconstructing the correct form of the target label.

Notwithstanding that growing up in a multilingual household is the global norm, differences between mono- and multilingual language development are not well understood: Only a handful of studies to date have considered multilingual (in particular, bilingual) lexical knowledge. Whether early bilingual lexical representations are more or less detailed / stable than monolingual ones given the inherent complexity of the multilingual language environment is an open empirical question. One can approach this question by determining how the mispronunciation detection skills of mono- and bilingual children differ, if at all. If bilingual children were more sensitive to phonological contrasts than monolinguals, it would enable building finer-grained lexical representations (and/or have better access thereto). In contrast, attenuated bilingual ability in mispronunciation detection could indicate weaker, less stable lexical representations, possibly due to relatively less experience with and exposure to each language in comparison to monolinguals.

Using the switch task, studies have found that both bilingual and monolingual 17- to 20-month olds dishabituated after the presentation of the phonetically different stimulus provided the speakers' language background matched theirs (i.e., bilingual infants learned more successfully from bilingual speakers and monolinguals from monolingual speakers) [12, 13, 14]. These findings show that both mono- and bilinguals are similarly attuned to phonetic detail and use such information to guide their learning of minimally different words. On the flip side, children from either a mono- or bilingual background may be disadvantaged when tested on non-matching stimuli, suggesting that bi- and monolingual infants are similarly constrained by the set of phonetic characteristics present in their environment [12, 13, 14].

In support of bilingual advantage, Mandarin-English bilinguals were found to be able to detect tone contrasts (e.g., a contrast between Tones 1 vs. 4) in newly learned words earlier than Mandarin monolinguals (12-13 vs. 17-18 months) [1]. Even with a non-tone language background, tonal sensitivity can be gained by bilinguals earlier than by monolinguals (11-12 vs. 17-18 months) [15]. To account for the bilingual success in these tasks, the authors cited cognitive adaptations that may accompany phonological conflict introduced by exposure to a tone- and non-tone language. This invites the possibility that, instead of being a source of confusion, under certain conditions the multilingual environment may in fact boost the identification of contrastive properties and thus the formation of word-object associations.

Similarly to [11], this study analyzed pupil dilation captured in a preferential looking paradigm (the eye tracking data is reported in [16]). In accord with the findings of past studies using pupillometry, mispronunciation was expected to increase the effort of recognizing the heard label and integrating it with the target image and the corresponding lexical entry, resulting in larger degrees of pupil dilation [9, 10, 11]. Critically for this study, given the existing literature on mono- vs. bilinguals' different abilities, the mispronunciation effect was predicted to be modulated by language background.

2. Method

Sixty two-year-old infants (1;11;3 – 2;2;1,M = 2;0;6, 27 boys) participated, 13 of which did not provide data analyzable by the eye tracker due to track loss, calibration error, and noncompliance. Infants were categorized as either Mandarin-English bilinguals (Mandarin exposure: 50-75%) or as Mandarin monolinguals (Mandarin exposure: $>$90%). Participants who did not fit those categories (N = 6) were excluded. Finally, those participants who did not reach a threshold of 50% of successful trials (those trials that contain pupil information from at least half the length of the trial) were excluded from further analyses, leaving 14 bilingual and 18 monolingual participants. On average, 81% of trials per participant were retained.

A total of 36 different target words (and 2 correctly produced practice words) were recorded by a Mandarin native speaker. All labels were monosyllabic nouns preceded by the exclamation 'Look!' in Mandarin Chinese. Auditory stimuli were normalized for amplitude and presented at 70 dB using Praat (version 5.3.63, [16]). Eighteen of these target words were correctly produced, 6 contained a vowel, 6 a consonant change (made to the word onset) and 6 a tone change. Each phonemic change constituted a single-feature deviation from the correct form. Out of the four lexical tones used by Mandarin Chinese (high level [Tone 1],rising [Tone 2], dipping [Tone 3], and falling [Tone 4]), three of which were used in this study. Tone mispronunciation included change between Tones 1 and 2, 2 and 4, and 1 and 4 (with the direction of change counterbalanced). Easily recognizable color pictures depicting a referent of the original word were converted such that all pictures were of a similar size (approximately 200 x 200 pixels displayed in a 320 x 320 pixel area). The areas of interest included the 450 x 450 pixel area around each picture.

In each trial, a pair of target and distractor images were simultaneously presented on each side of the screen on a white background for 2500 ms, the target depicting a familiar item and the distractor an unfamiliar item. After the center fixation phase, the same pair of pictures was shown for 2500 ms, accompanied by a Mandarin auditory label for the target image that was either correctly pronounced or mispronounced (consonant / tone / vowel change). Four versions of the task were created, each item occurring once in each version with the mispronunciation types counterbalanced across the four versions; children never saw the same picture or heard the same label more than once. Each participant was randomly assigned to one of the versions. Prior to the experiment, 2 practice items were presented in order to familiarize the child with the task. The practice items were not included in the analysis. Altogether, participants were presented with 20 correctly and 18 incorrectly pronounced items in each version of the experiment.

The study was conducted on a Tobii 60 XL eye tracker (version 3.2.1) coupled with a 24" LCD monitor used for the presentation of stimuli and the recording of infant eye gaze and an experimenter computer (Dell Optiplex 755). Children were seated 70–80 cm away from the screen. The auditory stimuli were presented via the in-built speakers on both sides of the LCD monitor.

3. Results

The prediction that language background modulated the mispronunciation effect was supported by the analyses. Linear mixed effects models were employed with random intercepts and slopes (estimates were chosen to optimize the log-likelihood criterion) [18]. `Mispronunciation type` (correct / consonant change / tone change / vowel change) and `language background` (bilingual / monolingual) were introduced as fixed effects, subjects and items as random effects, and corrected pupil size change (the mean of the 100 ms interval prior to the auditory label onset subtracted from the raw pupil value) was used as the outcome measure. `Mispronunciation type` was treatment-coded such that the correct condition was compared to all other conditions. Each intercept and slope fitted by the model was adjusted by the effect of `mispronunciation type` nested in participants [19]. The correlation term for `mispronunciation type` in the random effect structure was removed [20]. The most parsimonious model was chosen through comparisons using Likelihood Ratio Tests [21] using the `anova` function from the `stats` package [22]. Thus the interaction model was chosen over the main effects model ($\chi^2(3)$ = 8.54, $p < .036$). This model contained a significant `mispronunciation type` x `language background` interaction, driven by the correct vs. consonant and correct vs. vowel change contrasts in bilinguals (β_1 = 0.02, SE = 0.01, t = 1.98, β_2 = 0.03, SE = 0.01, t = 2.21, respectively, the correct vs. tone change contrast was not significant: t $<$ 1.23) and no significant contrasts in monolinguals (all ts $<$ 1.10).

Time-course analyses (post-hoc cluster-based permutation tests [23, 24]) were used to explore when significant differences emerged across the factor levels (4 levels of `mispronunciation type` and 2 levels of `language background`). First, individual paired sample t-tests found the significant ($p < .05$) t-values across the whole time frame. Second, clusters (e.g., contiguous significant t-values) were identified, for which a cluster-level t-value was given as the sum of all

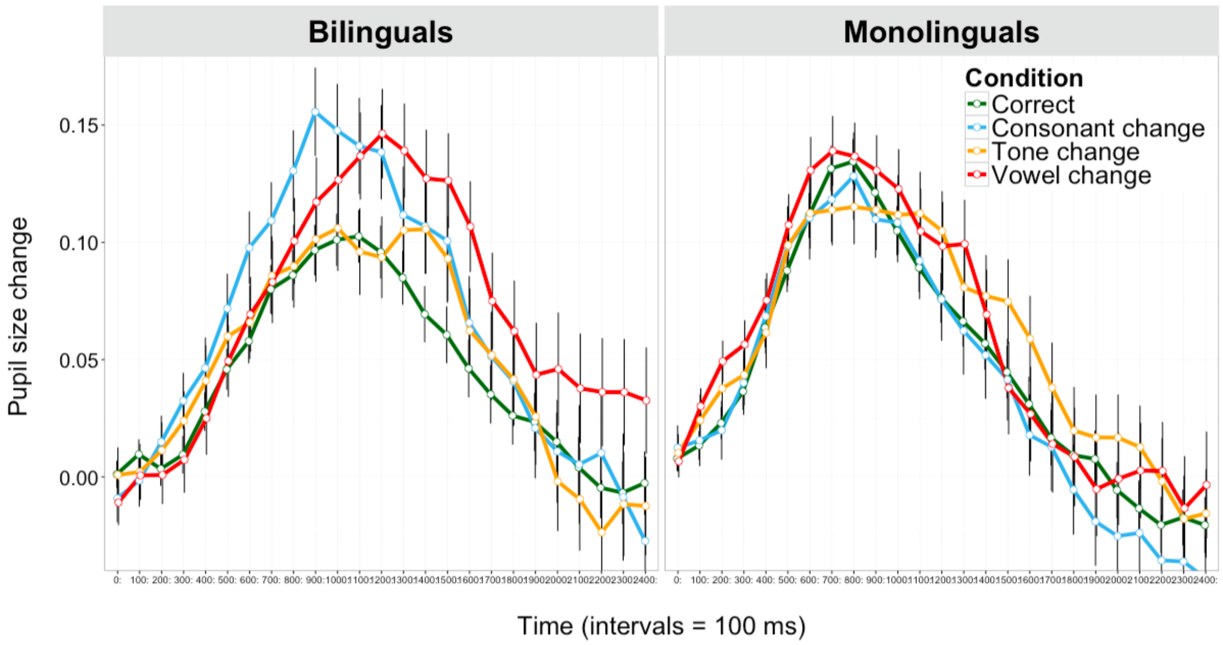


Figure 1: *Corrected pupil size change by mispronunciation type and language group (auditory label onset = 0 ms, error =* $95\%CI$*).*

single sample t-values within the cluster. Third, the significance of cluster-level t-values were assessed by generating Monte Carlo distributions (N = 2000) thereof and determining the probability of their occurrence given the distribution. Those clusters whose t statistic exceeded the threshold (t = 2.8, Bonferroni-corrected for multiple comparisons) were then tabulated for each contrast. With this method, significant clusters in the bilingual group were identified. Positive cluster-level t-values signal that all mispronounced conditions were associated with larger pupil dilation than the correct condition among bilinguals (lines 1–3 in Figure 1). No significant clusters in the monolingual group were found (c.f., lines 4–6 in Figure 1). Furthermore, comparisons across the bi- and monolinguals yielded significant clusters in all conditions (c.f., lines 7–9 in Figure 1). Positive cluster-level t-values indicate that monolinguals' pupils dilated significantly more than bilinguals', while negative values signal significant differences in the opposite direction. Thus in the consonant change condition, bilinguals exhibited a higher degree of pupil dilation than monolinguals and in the vowel and tone change conditions, both directions can be observed.

4. Discussion

This study supports previous research [1, 15] that has found a bilingual advantage in word recognition: The significance of the correct vs. mispronounced contrast in the bilingual and lack thereof in the monolingual group indicate that bilinguals exhibited more sensitivity to segmental change than their monolingual peers, which corroborates previous work showing bilingual children to be more sensitive to contrasts than monolinguals [1, 15]. Time-course post-hoc analyses confirmed those findings from linear mixed effects models and additionally found the correct vs. tone contrast to be significant in the bilingual group.

Time-course analyses revealed that hearing the correct form induced greater degrees of pupil dilation for monolingual vs. bilingual speakers. Contrastively, hearing consonant-change mispronunciation produced the opposite pattern and when processing the other mispronounced labels, the overall pupil dilation provided by the two language groups were comparable.

Table 1: *Summary table of time-course analyses (Interval = time interval in the naming phase,* $\sum t$ *= cluster-level t, p = p-value associated with cluster-level t, Corr. = correct label,* ΔC *= consonant change,* ΔV *= vowel change,* ΔT *= tone change).*

Contrasts	Interval	$\sum t$	p
Bilinguals			
Corr. vs. ΔC	800–1500	16.60	*
Corr. vs. ΔV	1100–1600	38.12	**
Corr. vs. ΔT	1400–1500	3.31	*
Monolinguals			
Corr. vs. ΔC	–	–	n.s.
Corr. vs. ΔV	–	–	n.s.
Corr. vs. ΔT	–	–	n.s.
Bi- vs. Monolinguals			
Corr.	200–900	14.83	*
ΔC	900–1800	−18.91	**
ΔV	100–900	18.55	**
ΔV	1200–1900	−11.86	*
ΔT	500-700	3.41	*
ΔT	1400-1500	−3.23	†

†: $p < .1$, *: $p < .05$, **: $p < .01$, n.s. = not significant

Considering pupil dilation to be a direct measure of cognitive effort [9, 10, 11], those findings can be interpreted such that bilinguals use fewer cognitive resources to activate the respective lexical representation than monolinguals. In a similar vein, bilinguals require more resources than monolinguals to activate the lexical entry when the label is only a partial match, showing greater specificity (or less flexibility) than monolinguals, at least with consonants.

Finding no evidence for the mispronunciation effect from the monolingual pupillary response is intriguing because it was demonstrated with 30-month-old monolinguals in previous research (i.e., based on their pupil size changes, monolingual children reliably differentiated between correct and mispronounced labels) [9, 10, 11]. The apparent discrepancy of monolingual performance versus the present null-result may be due to inherent differences with respect to manipulation, paradigm, and participant age. It is challenging to compare the study with previous research that explored the effect of featural distance and thus manipulated the number of feature mispronunciations produced by a monolingual speaker [10, 11] (vs. type of mispronunciations produced by a bilingual speaker), using single-picture pupillometry paradigms [9, 10] (vs. the preferential looking paradigm) with 30-month-olds [9, 10, 11] (vs. 24-month-olds). Of these potential reasons for discrepancy, age seems the least likely as younger children have been shown to be sensitive to mispronunciations with other methodologies [2, 3, 4, 5].

It is worth noting that in this study, the attrition rate was high due to calibration error and track loss, barring eye and pupil tracking for 13 children. Once more data are collected, further analyses can be carried out that allow for the restriction of the analysis window to target looks (similarly to [11]). Nevertheless, results at this stage suggest that bilinguals' pupillary response was more sensitive to phonemic contrasts than monolinguals' (a finding consistent with those of [1, 15]).

5. Conclusions

The present study is the first to offer evidence that children's mispronunciation detection skills – as measured by pupillary reactions – are affected by language background. Results show that for bilinguals, mispronounced labels yielded an increase in pupil dilation in comparison to correct labels, whereas for monolinguals, no significant differences in pupil dilation were observed. Furthermore, reliable differences were recorded between mono- and bilinguals at two levels of mispronunciation type. Correct labels were associated with smaller degrees of pupil dilation – while consonant change with larger degrees of pupil dilation – in the bilingual vs. in the monolingual group.

Following past research [9, 10, 11], we interpret changes in the pupillary response as an indicator of resource consumption. As such, bilingual children seem to require less effort to link the correct label to its corresponding representation than monolingual children, suggesting more economical processing. On the other hand, establishing such a link when the label has a consonant change seems to be more demanding for bilingual than monolingual children. Methodologically, this study contributes to the growing body of literature demonstrating pupillometry to be a viable - dynamic and gradient - tool to study early word recognition [9, 10, 11], hence contributing to our understanding of mono- and bilingual developmental trajectories.

6. Acknowledgments

This research was supported by a Tier 1 grant to LS (FY2013-FRC2-009). KT and BH was supported by EMJD 520101-1-2011-1-DE-ERA. We would like to thank Wong Liang Hui and Felicia L.S. Poh for recruiting and testing all of the participants.

7. References

[1] Singh, L., Poh, F. L., and Fu, C. S. (2016). Limits on Monolingualism? A Comparison of Monolingual and Bilingual Infants Abilities to Integrate Lexical Tone in Novel Word Learning. *Frontiers in Psychology, 7:* 667.

[2] Swingley, D. (2005). 11-month-olds' knowledge of how familiar words sound. *Developmental Science, 8*(5), 432-443.

[3] Yoshida, K. A., Fennell, C. T., Swingley, D., and Werker, J. F. (2009). Fourteen-month-old infants learn similar-sounding words. *Developmental Science, 8*, 12(3), 412-418.

[4] Swingley, D., and Aslin, R. N. (2000). Spoken word recognition and lexical representation in very young children. *Developmental Science, 8*(2), 147-166.

[5] Mani, N., Mills, D. L., and Plunkett, K. (2012). Vowels in early words: an event-related potential study. *Developmental Science, 15*(1), 2-11.

[6] Karatekin, C. (2007). Eye-tracking studies of normative and atypical development. *Developmental Review, 27*(3):283348.

[7] Hochmann, J.-R. and Papeo, L. (2014). The invariance problem in infancy: A pupillometry study. *Psychological Science, 25*(11):2038-46.

[8] Kuipers, J. and Thierry, G. (2011). N400 amplitude reduction correlates with an increase in pupil size. *Frontiers in Human Neuroscience, 5.*

[9] Fritzsche, T., and Höhle, B. (2015). Phonological and lexical mismatch detection in 30-month-olds and adults measured by pupillometry. In The *Proceedings of the ICPhS, (18),* 03-39.

[10] Tamási, K., McKean, C., Gafos, A., Fritzsche, T., & Höhle, B. (in press). Pupillometry registers toddlers' sensitivity to degrees of mispronunciation. *Journal of Experimental Child Psychology*.

[11] Tamási, K., McKean, C., Gafos, A., & Höhle, B. (2016). *Children's sensitivity to degrees of mispronunciation: Enriching the preferential looking paradigm with pupillometry.* Manuscript in preparation.

[12] Fennell, C., and Byers-Heinlein, K. (2014). You sound like Mommy: Bilingual and monolingual infants learn words best from speakers typical of their language environments. International Journal of Behavioral Development, 38(4), 309-316.

[13] Fennell, C. T., Byers-Heinlein, K., and Werker, J. F. (2007). Using speech sounds to guide word learning: The case of bilingual infants. *Child Development, 78*(5), 1510-1525.

[14] Mattock, K., Polka, L., Rvachew, S., and Krehm, M. (2010). The first steps in word learning are easier when the shoes fit: Comparing monolingual and bilingual infants. *Developmental Science, 13*(1), 229-243.

[15] Liu, L., Kager, R. (2016). Perception of Tones by Bilingual Infants Learning Non-Tone Languages. *Bilingualism: Language and Cognition, 1-15.*

[16] Wewalaarachchi, T.D., Wong, L.H. and Singh, L. (2016). *Sensitivity to phonological detail in bilingual and monolingual infants: Evidence from familiar word recognition.* Manuscript submitted for publication.

[17] Boersma, P., and Weenink, D. (2013). *Praat: doing phonetics by computer [Computer program] (Version 5.3.51).* Retrieved from http://www.praat.org/

[18] Bates, D. (2005). Fitting linear mixed models in R. *R news, 5*(1), 27-30.

[19] Barr, D. J., Levy, R., Scheepers, C., and Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language, 68*(3):255278.

[20] Jaeger, T. F., Graff, P., Croft, W., and Pontillo, D. (2011). Mixed effect models for genetic and areal dependencies in linguistic typology. *Linguistic Typology, 15*(2):281 320.

[21] Pinheiro, J., Bates, D., DebRoy, S., and Sarkar, D. (2007). Linear and nonlinear mixed effects models. R-Package-version, 3:57.

[22] R Core Team (2014). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

[23] Maris, E., and Oostenveld, R. (2007). Nonparametric statistical testing of EEG-and MEG-data. *Journal of Neuroscience Methods, 164*(1), 177-190.

[24] Dink, J. W., and Ferguson, B. F. (2015). *eyetrackingR: An R Library for Eye-tracking Data Analysis.* Retrieved from http://www.eyetrackingr.com.

Sensitivity to Vowel, Consonant and Tone Variation in Early Childhood

Thilanga D. Wewalaarachchi and Leher Singh

National University of Singapore

thilanga.d.w@u.nus.edu, psyls@nus.edu.sg

Abstract

Children have to possess robust phonological representations in order to become proficient language users. Mandarin Chinese learners' spoken word recognition abilities were investigated via an eye-tracking paradigm. Kindergarteners (5-6 years) were presented with correct pronunciations and mispronunciations involving a vowel, consonant or tone substitution. A robust ability to process correct pronunciations relative to mispronunciations was found. However, sensitivities to vowel, consonant and tone mispronunciations were not comparable. Results point to a divergence in the phonological sensitivities of segments and tones. Findings are discussed in terms of the properties and function of these sources of variation in tone languages.

Index Terms: Language Acquisition, Phonological Development, Lexical Tone

1. Introduction

A significant challenge in early language learning is the development of language-specific phonological sensitivities. Children have to learn to reject incorrect pronunciations of words and only accept pronunciations that fall within the prescribed boundaries of each phonetic category represented in the word. To date, most of the prior work pertaining to childrens' phonological development has been with intonation language learners such as learners of English [e.g. 1, 2]. Intonation languages employ two sources of phonemic variation in distinguishing lexical meaning: vowels and consonants. However, most children learn languages that employ an additional source of phonemic variation: lexical tone [3]. A disproportionate focus in research is problematic as theories pertaining to the development of phonological sensitivities are then limited to findings obtained from studying language typologies that form a statistical minority. The present study aims to compare vowel, consonant and lexical tone sensitivity in a crucial area of language development: spoken word recognition.

One way to investigate phonological development is to observe children's visual responses to correct pronunciations and incorrect pronunciations of familiar words (mispronunciation paradigm). In this way, the cost of word recognition that is associated with the substitution of each source of phonemic variation can be directly compared. It is not yet clear whether vowels, tones and consonants are equally well specified in the early childhood lexicon, which serves as the purpose of this study. In the recent years, work with adults have highlighted that the relative constraints that vowels, consonants and lexical tone place on word recognition depend on the type of language in question [4, 5]. In some intonation languages, it has been suggested that consonants are more important than vowels at the level of the lexicon [5]. However, a bias to prioritize consonant information is not language universal. This is not the case in tone languages, where vowel information constrains word recognition more than consonants [4]. Likewise, patterns of phonological development for children learning intonation versus tone languages are likely to be unique.

A possible reason that vowels, consonants and tones exert different effects on word recognition is that they are compositionally distinct [6]. Lexical tones differ from vowels and consonants in that tone results in syllable-level or suprasegmental changes, while vowels and consonants result in segmental changes [6]. Other types of suprasegmental cues in language include lexical stress and intonation. In addition, lexical tone is defined primarily by fundamental frequency or pitch, while vowels and consonants can be defined in terms of formant properties [6]. Vowel features are determined by tongue position and lip rounding while consonant features are characterized by phonation, manner and place of articulation. Although the acoustics of segmental cues and tonal cues differ, recent evidence suggests that tonal information is not processed at a protracted rate relative to segmental information [7].

Prior research demonstrates that toddlers (2 year-olds) are initially equally sensitive to mispronunciations involving vowels, consonants and tones [8]. However, at a later stage, younger pre-schoolers (2.5-3 year-olds) are more sensitive to tone mispronunciations and less so to mispronunciations involving vowels and consonants [9]. Interestingly, older pre-schoolers (4-5 year-olds) are more sensitive to mispronunciations involving vowels and consonants and less so to tone mispronunciations [9]. However, given that Mandarin learning pre-schoolers only begin to reconcile intonation functions of tone (question and statement distinction) between the ages of 4 and 5 [10], it is possible that mispronunciation effects for tone temporarily weaken because older pre-schoolers are in the midst of negotiating the complex differentiation of tonal and intonational functions during this stage of development. The preschool years are a period of aggressive language development [1]. It is possible that after a temporary period of attenuation to tone due to the functional differentiation of pitch during the preschool years, tone sensitivity may strengthen in school-aged children. The goal of the present study is to build upon prior research [9] to determine whether the previously observed attenuation in tone sensitivity extends through the next year. The present study investigates the relative impact of vowel, consonant and tone identity on lexical disambiguation in Mandarin Chinese learning kindergarteners.

2. Method

2.1. Participants

Eighteen native learners of Mandarin Chinese were sampled for this study (nine boys). Participants were between the ages of five and six (*M*= 70.94months, *SD*= 3.26months), and had no known disabilities or developmental delays.

2.2. Stimuli

Eighteen concrete, imageable words that were judged to be familiar to children aged 5 to 6 were chosen as test stimuli. A post-experimental vocabulary test was conducted to ensure that these early-acquired words were indeed familiar to individual participants. During this receptive vocabulary test, test stimuli were paired, and participants were asked "哪一个是 [target]?" (English translation: which is the [target]?). Children performed at 100% accuracy.

Test stimuli belonged to four trial types: correct pronunciations, vowel, consonant, and tone mispronunciations (please see Table 1 for sample stimuli). The higher number of correct pronunciation trials relative to the number of individual mispronunciation trials served to sustain participants' attention to the experimental task [9]. All mispronunciations resulted in non-words.

Trial Type		Test Stimuli (Translation)
Practice		Flower
		Tree
		Door
Correct Pronunciation		Pig
		Shoe
		Egg
		Ruler
		Deer
		T-shirt
		Watch
		Fork
		Dress
Tone Mispronunciation	2 to 1	Cow
	4 to 2	Rice
	4 to 1	Noodles
Vowel Mispronunciation	Backness	Paper
	Height	Trousers
	Roundedness	Chicken
Consonant Mispronunciation	Place	Pen
	Manner	Car
	Aspiration	Ball

Table 1. *Sample Stimuli List*

During the experiment, test stimuli were presented in sentence final position, with the carrier phrase "你看，那是 [target]" (English translation: Look, that is a [target]). A female adult native Mandarin Chinese speaker recorded all auditory stimuli in a child-directed manner.

Visual stimuli comprised of photographed target objects (test stimuli) and distractor objects. Novel distractors were chosen to serve as a potential alternative referent in mispronunciation trials [9]. Positions of familiar target and novel distractor were randomized and counterbalanced.

2.3. Procedure

A preferential looking paradigm was employed to investigate spoken word recognition. All participants were tested in a quiet room with a caregiver present. A 17-inch Macintosh computer was placed at participants' eye-level at a distance of 30cm. Auditory stimuli were played via external speakers at a conversational level of 70dB. The experiment comprised of 3 practice trials and 18 test trials. The purpose of these practice trials was to initiate children to the paradigm and so data from the first three trials were not analyzed.

Each trial comprised of two phases of equal duration. During the pre-naming phase, participants heard the neutral directive "Look, that is a". During the post-naming phase, participants heard the test stimulus, which was synchronized to begin at the 2500ms mark. The static visual display (familiar target and novel distractor) remained on screen for the entire duration of the trial (please see Figure 1 for sample trial structure).

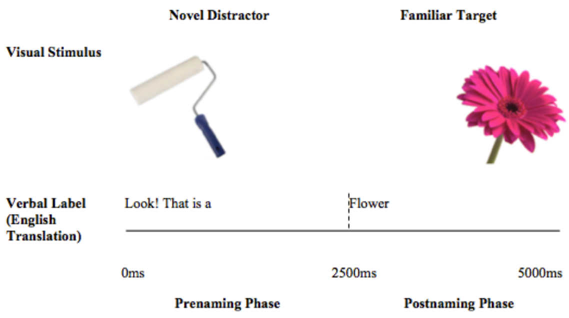


Figure 1. *A Schematic Diagram of Trial Structure*

3. Results

Participants' eye movements were coded offline, frame-by-frame at a rate of 30 frames per second. For each frame, the coder determined if the participant was looking left, right or elsewhere. As per convention, the time window used for statistical analysis was from 200ms to 1200ms after the onset of the test stimulus [2]. Figure 2 depicts the time course of word recognition for this duration.

Analyses comprise of investigations pertaining to the time course of spoken word recognition. Unlike analyses of naming effects (e.g. [9]), time course analyses provide insight on the temporal dynamics of lexical disambiguation, reflecting efficiency of online processing that occurs in real-time as the sentence unfolds. Time course analyses have the potential to reveal a fine-grained view into the constraints on spoken word recognition [11]. This is particularly relevant in the present study to guide conclusions about how tonal information is processed relative to segmental information (vowels and consonants).

To allow for time required to program an eye-movement as a result of hearing the auditory label, a 200ms time allowance was afforded [2]. Should the auditory label correspond to a correct pronunciation, a sustained fixation to the target object is expected. On the other hand, should the auditory label correspond to a mispronunciation, a decrease in fixation to the target object and a sustained fixation to the distractor object is

Table of Contents for this manuscript

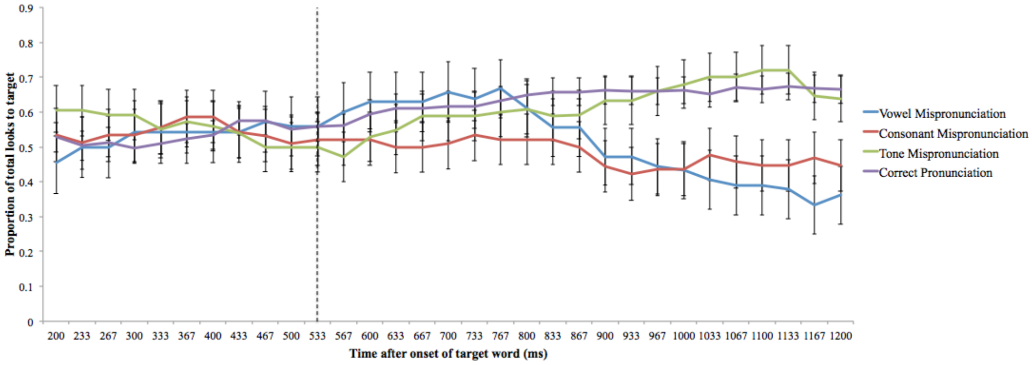


Figure 2. *Proportion of total looks to target during post-naming (Error bars reflect SEM). Dashed line indicates average offset of test stimuli.*

expected. For mispronunciation trials, higher rates of 'false-alarm' indicated by greater fixation to the target object are indicative of less efficient processing. To allow for analysis of fixation patterns by epoch, gaze data was bundled into five 200ms time intervals e.g., 200-400ms, 400-600ms … 1000-1200ms [2].

A repeated-measures ANOVA with trial type as independent factor, time interval as between subject factors, and PTL as dependent measure was conducted. Trial type comprised of four levels: correct pronunciation, vowel, consonant, and tone mispronunciation. There was a main effect of trial type, $F(1,25) = 133.47$, $p < .0001$ (partial eta^2 = .84), revealing that time course of word recognition statistically differed depending on trial type. In addition, there was a main effect of time interval, $F(4,25) = 12.13$, $p < .0001$ (partial eta^2 = .1), revealing that time course of word recognition statistically differed depending on time interval. Lastly, there was a significant two-way interaction between trial type and time interval, $F(4,25) = 60.61$, $p < .0001$ (partial eta^2 = .91), revealing that the differential time course patterns observed for each trial type varied systematically depending on time interval.

Post-hoc comparisons were computed to statistically examine these effects. For each time interval, post hoc comparisons were computed between trial types. Pairwise comparisons (with Bonferroni corrections) revealed greater fixations to target for correct pronunciations relative to tone mispronunciations for zero continuous time intervals. In contrast, pairwise comparisons (with Bonferroni corrections) revealed greater fixations to target for correct pronunciations relative to vowel mispronunciations for four continuous time intervals (400-600ms, 600-800ms, 800-1000ms and 1000-1200ms). Likewise, pairwise comparisons (with Bonferroni corrections) revealed greater fixations to target for correct pronunciations relative to consonant mispronunciations for four continuous time intervals (400-600ms, 600-800ms, 800-1000ms and 1000-1200ms). Contrasting both types of segmental mispronunciations, pairwise comparisons (with Bonferroni corrections) revealed neither vowel nor consonant mispronunciation had greater fixation to target for any continuous time intervals.

In summary, time course analyses revealed that sources of phonemic variation exert differential constrains on spoken word recognition. Tone information was found to affect word recognition to a lesser degree as compared to segmental information. In addition, vowels and consonants affected word recognition to the same degree.

4. Discussion

In the present study, kindergarteners who were native speakers of Mandarin Chinese were tested on their sensitivity to vowel, consonant and tone substitutions within a spoken word recognition task. Participants demonstrated a robust ability to process correct pronunciations relative to mispronunciations. When processing correct pronunciations, participants showed steady and sustained fixations toward the target object. Fewer fixations toward the target object were observed when processing vowel, consonant and tone mispronunciations. However, sensitivities to vowel, consonant and tone mispronunciations were not comparable. There was a striking divergence in the time course of spoken word recognition when processing of segmental information versus tone information. Sensitivities to vowels and consonants were comparable: vowels did not constrain word recognition more than consonants. However, sensitivities to tones were weaker than sensitivities to segments.

The dissociation observed between segmental information and tonal information complements prior work with children [9] and adult Mandarin Chinese speakers [12]. One postulate that has been put forth to explain this finding is the information load hypothesis [4, 12]. In essence, the higher the likelihood that a linguistic signal appears in speech the lower the information value it confers [13]. This probability is related to inventory size. In Mandarin Chinese, the size of the tone inventory is very small relative to the size of the segment inventories [12]. Therefore, it is possible that tones due to its low information value are assigned lower priority in conditions of limited resources. A thorough statistical account contrasting various tone languages that have a different ratio of tones to segments awaits further investigation.

It should be noted that work with adult tone language speakers has typically observed two distinct patterns of tone sensitivity depending on the type of task and nature of stimuli used. Under experimental conditions that allow for pre-activation of the target word: the use of pictures [7] or highly predictable sequence of words [14], tone sensitivity is comparable to segmental sensitivity. On the other hand, under experimental conditions that are more ambiguous and/or challenging such as lexical reconstruction [4] and speeded classification [12], tone sensitivity is weaker relative to segmental sensitivity. It could be that during the period of early childhood, children are still trying to reconcile how much priority ought to be attributed to tone information.

Taken together with findings from [9], the present study suggests that weak sensitivity to tone information relative to segments is not temporary. Rather, it appears that native learners of Mandarin Chinese go through a period of prolonged weak sensitivity to tone mispronunciations through the preschool and kindergarten years. The observation of a sustained weak mispronunciation effect for a source of phonemic variation is surprising given maturation effects previously observed with intonation language learners processing segments in word recognition [1].

The present study provides further evidence to the difficulty of generalizing findings obtained from intonation language learners to all language learners and attest to the promise of investigating sensitivity to lexical tone – in conjunction with vowels and consonants – to advance a comprehensive view of the developing mental lexicon. It appears that the developmental trajectory of lexical tone acquisition is distinct to that of segmental acquisition.

5. Funding

Ministry of Education Tier 1 Academic Research Fund grant (FY2013-FRC2-009) and a start-up grant from the National University of Singapore to LS.

6. Acknowledgements

We are grateful to Felicia Woo for assistance with stimulus recording and Hwee Hwee Goh for assistance with experimental design.

7. References

[1] Havy, M., Bertoncini, J., & Nazzi, T. (2011). Word learning and phonetic processing in preschool-age children. *Journal of experimental child psychology, 108*(1), 25-43.

[2] Creel, S. C. (2012). Phonological similarity and mutual exclusivity: on-line recognition of atypical pronunciations in 3–5-year-olds. *Developmental Science, 15*(5), 697-713.

[3] Yip, M. (2002). *Tone*. Cambridge, UK: Cambridge University Press.

[4] Wiener, S., & Turnbull, R. (2016). Constraints of tones, vowels and consonants on lexical selection in Mandarin Chinese. *Language and speech*, *59*(1), 59-82.

[5] Van Ooijen, B. (1996). Vowel mutability and lexical selection in English: Evidence from a word reconstruction task. *Memory & Cognition*, *24*(5), 573-583.

[6] Singh, L., & Fu, C. S. (2016). A new view of language development: the acquisition of lexical tone. *Child development, 87*(3), 834-854.

[7] Malins, J. G., & Joanisse, M. F. (2010). The roles of tonal and segmental information in Mandarin spoken word recognition: An eyetracking study. *Journal of Memory and Language*, *62*(4), 407-420.

[8] Wong, L. H, & Singh, L. (2015) Sensitivity to phonological detail in bilingual and monolingual infants. Poster presented at Workshop on Infant Speech Perception: WISP. Sydney, NSW

[9] Singh, L., Goh, H. H., & Wewalaarachchi, T. D. (2015). Spoken word recognition in early childhood: Comparative effects of vowel, consonant and lexical tone variation. *Cognition*, 142, 1-11.

[10] Singh, L., & Chee, M. (2016). Rise and fall: Effects of tone and intonation on spoken word recognition in early childhood. *Journal of Phonetics*, *55*, 109-118.

[11] Swingley, D. (2011). The looking-while-listening procedure. *Guide to Research Methods in Child Language*.

[12] Tong, Y., Francis, A. L., & Gandour, J. T. (2008). Processing dependencies between segmental and suprasegmental features in Mandarin Chinese. *Language and Cognitive Processes, 23*(5), 689-708.

[13] Garner, W. R. (1988). The contribution of information theory to psychology. In W. Hirst (Ed.), *The making of cognitive science: Essays in honour of George A. Miller*, 19-35. Cambridge, UK: Cambridge University Press.

[14] Liu, S., & Samuel, A. G. (2007). The role of Mandarin lexical tones in lexical access under different contextual conditions. *Language and Cognitive Processes*, *22*(4), 566-594.

The effects of allomorphic variation on children's acquisition of plural morphology

Benjamin Davies, Nan Xu Rattanasone, Katherine Demuth

Department of Linguistics, Center for Language Sciences, Macquarie University.
ben.davies@mq.edu.au, nan.xu@mq.edu.au, katherine.demuth@mq.edu.au

Abstract

While children begin producing plural words in their natural speech from around two years, it is unclear when they acquire a full understanding of plural morphology. Two intermodal preferential looking (IPL) experiments using eye-tracking examined the effect of allomorphic variation on children's acquisition of English plural morphology. Experiment One looked at 24-month-olds' understanding of segmental plural allomorphs /s/ and /z/. It was found that, while children at this age did not understand the number condition of nonce CVC singular words, or for nonce plural words inflected with voiced plural /z/, they did demonstrate understanding of the voiceless plural allomorph /s/. Experiment Two then tested 36-month-olds' understanding of the syllabic plural /əz/, finding that at this age children are able to demonstrate understanding of CVC singular nonce words, and of nonce words inflected with syllabic plural /əz/. These results add to our understanding of how allomorphic variation affects children's acquisition of nominal plural morphology.

Index Terms: language acquisition, morphology, plural

1. Introduction

English-acquiring children begin using plural words in their day-to-day speech from around the age of two [1, 2, 3], and non-linguistic research shows that a semantic contrast of *one* vs. *more-than-one* has developed by around this age [4, 5]. It is not clear, however, when children develop an understanding of nominal plural morphology. That is, when they understand that a word such as *cats* has an internal structure comprising of a root morpheme, /kæt/, denoting its referent, and a plural morpheme, /s/, denoting its number (i.e., singular or plural). It is also not known what effect allomorphic variation has on children's acquisition of plural morphology. Children must learn that form of the plural morpheme is dependent on the phonological properties of the root morpheme. There are three plural allomorphs in English: voiceless plural /s/, which occurs when the final segment of the root morpheme is a voiceless consonant (e.g., *cats* /kæt**s**/); voiced plural /z/, which occurs when the final segment is voiced (e.g., *dogs* /dɔg**z**/); and the syllabic plural /əz/, which attaches to strident fricative (e.g., buses /bɐs**əz/**). Understanding plural morphology and allomorphy not only allows children to inflect words into plural forms, but gives them the ability to readily identify the number condition of newly heard words. Children with an understanding of plural morphology and allomorphy would readily be able to identify *peas* /piːz/ as being plural, and *piece* /piːs/ as being singular, even without explicitly knowing the meaning of those words.

One research method that has been used to explore children's understanding of plural morphology and allomorphic variation has been the wug task [6]. In a wug task, children are presented with a picture of an unknown animal and a nonce label, such as *wug*. Next, they are presented with a picture depicting multiple of that same animal, and asked what they see. An understanding of plural morphology should allow children to inflect that nonce label into a previously unheard plural form, in this case, *wugs*. However, studies suggest that, while children have some fledgling ability at two years of age [7], they still lack a solid grasp of plural morphology even at the age of seven [6]. Results also suggest that allomorphic variation contributes to children's difficulty, as their performance with the syllabic plural /əz/ is especially poor compared to its segmental counterparts, voiceless /s/ and voiced /z/ [6].

Because wug tasks necessitate the production of segments that are perhaps difficult for children to pronounce, an alternative and potentially less demanding research method used to explore young children's linguistic knowledge is the Intermodal Preferential Looking (IPL) paradigm [8]. In a typical IPL study, children are presented with two pictures, one a target, the other a distractor. They are then played an auditory stimulus. Looking preferences between the two pictures are compared before and after hearing the stimulus. If children's looking preferences change towards the target picture, this is seen as evidence of understanding the task at hand.

One such IPL study examined children's understanding of plural morphology in task which shared some similarities to a wug task [9]. Children were presented with two pictures, one depicting a solitary object (singular picture) and the other depicting multiple unknown objects (plural picture). Children were then told to "Look at the [nonce-word]". The nonce word was either inflected for plural or singular. The study found that, when children's only cue was the presence/absence of the plural morpheme, 36-month-olds shifted their looking preference towards the target picture, but 24-month-olds did not. That is, 36-month-olds showed understanding of plural morphology, but 24-month-olds did not.

That study, however, was not designed to look at allomorphic variation, and tested children with an unequal mix of plural allomorphic variants /s/, /z/ and /əz/. Closer scrutiny, furthermore, revealed two points of interest in the results. Firstly, while 24-month-old participants appeared not to understand voiced plural /z/ or syllabic plural /əz/, they did show a potential sensitivity to voiceless plural /s/. However, this result was only approaching significance. Secondly, while 36-month-olds showed understanding of plural morphology overall, closer analysis raises doubts over their understanding of syllabic plural /əz/; unlike segmental plurals /s/ and /z/, post hoc analyses did not reveal /əz/ to be significantly above chance. This is interesting, as not only is the syllabic plural /əz/ the least frequent plural allomorph [1], it is also the one that has proven to be the most difficult in wug tasks [6].

Furthermore, it is unclear whether these results reveal that children have an understanding of plural morphology *per se*, as the auditory stimuli used in the task make it unclear as to whether children interpreted the plural-inflected nonce words as *root+plural morpheme*, or whether they simply identified fricative-final words as being plural.

Therefore, two IPL experiments were carried out to further probe children's understanding of nominal plural morphology, and to explicitly look at the effects of allomorphic variation. Experiment One tested 24-month-olds' understanding of segmental plural allomorphs /s/ and /z/ in phonologically simple CVC or CVCs/z contexts. While it has been shown that plural /z/ is more frequent in children's input [1], plural /s/ has been shown to be longer in duration and therefore more perceptually salient [10]. It was predicted that children would demonstrate some understanding of at least voiceless plural /s/, if not also voiced plural /z/.

Experiment Two then tested 36-month-olds' understanding of syllabic plural /əz/. Given that it is the most perceptually salient allomorphic variant, it was predicted that the children would be able demonstrate understanding of syllabic plural /əz/. Experiment two also tested on whether children would correctly identify the number condition of fricative-final singular nonce words, in CVs and CVz forms. Despite being fricative final, these words must be singular, as they cannot be decomposed into *root+plural morpheme*. It was not known whether children would identify these words as singular, or interpret them as plural.

2. Method

2.1. Experiment One

2.1.1. *Participants (Ex 1)*

Nineteen 24-month-old children participated (7 girls, 12 boys; Mean = 24 months; Range = 23 to 25 months). An additional 12 children were excluded for failure to return a sufficient number of trials (minimum 2 each of plural /s/, /z/ and singular trials), due to fussiness, inattention, or poor eye-tracking.

2.1.2. *Auditory Stimuli (Ex 1)*

Stimuli were produced by a female speaker of Australian English, using child-directed speech. Auditory stimuli included 12 nonce words, recorded as both CVC singular and CVCs/z plural conditions (see Table 1).

Table 1: *Experiment One Nonce Words*

	Singular	Plural
Voiceless /s/	mip	mips
	tep	teps
	gop	gops
	nep	neps
	gip	gips
	dup	dups
Voiced /z/	kib	kibz
	gub	gubz
	pog	pogz
	nug	nugz
	deg	degz
	tig	tigz

Audio were recorded as complete utterances with the carrier phrases "*look at the...*" and "*find the...*"

2.1.3. *Visual Stimuli (Ex 1)*

Visual stimuli were 16 unknown animals depicted with happy faces and closed eyes. The eyes were closed as it was thought open eyes would create the impression of being stared at from the plural picture, thus making it more visually attractive. Each animal was made as both a one-animal (singular) picture and a five-animal (plural) picture. Two versions were created so that no child saw both the singular and plural forms of the same animal.

2.1.4. *Procedure (Ex 1)*

Children sat on their parent's lap in a darkened room in front of a widescreen monitor. During each trial, two pictures (one singular, one plural) were displayed side-by-side for 5 seconds. A looming red ball in the center of a black screen then replaced the pictures for 1 second. The audio stimulus was then played over a black screen. The nonce word presented was either singular (e.g., *gop, deg*), inflected with voiceless plural /s/ (e.g., *gops*), or inflected with voiced plural /z/ (e.g., *degz*). The pictures then returned for 4 seconds. Children's looking behavior was recorded using a Tobii x120 eye tracker.

2.2. Experiment Two

2.2.1. *Participants (Ex 2)*

Twenty 36-month-old children participated (8 girls, 12 boys; Mean = 36 months; Range = 35 to 36 months). An additional 6 children were excluded due to fussiness and inattention.

2.2.2. *Auditory Stimuli (Ex 2)*

Auditory stimuli were 12 monosyllabic, fricative-final novel word stems (six /s/-final and six /z/-final). Each word stem was recorded as both a CVC singular word and a CVCəz plural-inflected word (see Table 2). Only short vowels were used to ensure CVC words were singular (as long vowels in CVz contexts can be both plural and singular e.g., cheese, fleas).

Table 2: *Experiment Two Nonce Words*

Singular	Plural
bess	besses
dass	dasses
dozz	dozzes
giss	gisses
gozz	gozzes
kazz	kazzes
koss	kosses
nass	nasses
nizz	nizzes
pezz	pezzes
poss	posses
tizz	tizzes

Auditory stimuli were recorded as complete utterances with the carrier phrase "*find the...*"

2.2.3. *Visual Stimuli (Ex 2)*

Similar to Experiment One, visual stimuli consisted of 16 unknown animals. However, visual stimuli for Experiment Two had an additional dancing animation (see below). Each animal had both a one-animal (singular) picture and a five-animal (plural) depiction.

2.2.4. *Procedure (Ex 2)*

In order to maintain children's attention and minimize participant exclusions, slight modifications were made to the procedure employed in Experiment One. During each trial, the two pictures (singular and plural) were displayed side-by-side for 4 seconds. The looming red ball replaced the pictures for 1 second. The audio stimulus was played over a black screen. The nonce word was either singular (e.g., *koss*, *nizz*), or inflected with segmental plural /əz/ (e.g., *kosses, nizzes*). The pictures then returned for 3 seconds. After 3 seconds the target picture danced for 1.5 seconds to a happy tune. The dancing was added to help maintain children's interest throughout the task.

3. Results

Difference Scores were used as the dependent measures for both experiment one and two. A difference score is calculated on a trial-by-trial basis, and is a measure of how much a child's looking preference shifts towards the target picture after hearing the audio stimulus. In order to calculate a difference score, children's looking preference towards the target picture is calculated for both before and after hearing the auditory stimulus (the dancing phase in Experiment Two was excluded from the analysis). Looking proportions are calculated by dividing the total fixation duration of the target picture by the sum total fixation durations recorded for both the target and distractor picture. Any time spent not looking at either picture was therefore excluded from the calculation. Difference scores were then calculated by subtracting the proportion looking to target pre-auditory stimulus from that of post-auditory stimulus. This figure was then multiplied by one hundred to gain a percentage. Positive difference scores indicate a child's preference shifted towards the target picture after hearing the audio stimulus, and vice versa for a negative shift.

3.1. Experiment One

With alpha set to 0.05, planned *t*-tests were first carried out on the 24-month-olds' difference scores for singular and plural (/s/ and /z/ collapsed) trials. Neither singular nor plural were found to be significantly above chance. Next, the plural allomorphs /s/ and /z/ were compared to chance. The voiceless plural allomorph /s/ was found to be significantly above chance (*t*(18) = 2.33, *p*=0.03), but the voiced plural /z/, was not different from chance (figure 1).

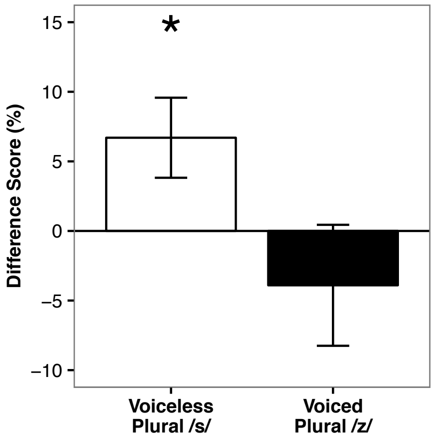


Figure 1: *24-month -olds' difference s cores for voiceless plural /s/ and voiced plural /z/. Error bars 1 SE. *p=.03*

3.2. Experiment Two

With alpha set to 0.05, planned *t*-tests were carried out comparing the singular and syllabic plural /əz/ difference scores to chance. For children aged 36-months, both singular and plural were significantly above chance (singular: *t*(19)=3.50, *p*<0.01), and plural: (*t*(19)=3.05, *p*<0.01) (figure 2).

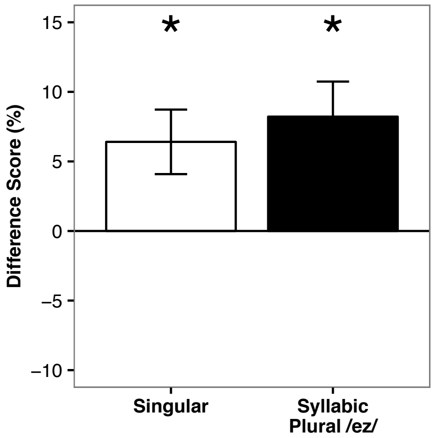


Figure 2: *36-month -olds' difference scores for singular and syllabic plural / əz/. Error bars 1 SE. *p<.01*

4. Discussion

Two IPL studies were carried out to examine whether allomorphic variation affects children's acquisition of plural morphology. Children aged 24 months were tested on their understanding of segmental plural allomorphs /s/ and /z/. Children aged 36 months were tested on their understanding of the syllabic plural allomorph /əz/, and whether they would interpret CVs and CVz nonce words as singular.

The results show that 24-month-olds understood the number condition of a never-before-heard nonce word so long as it was inflected with the voiceless plural allomorph /s/. Children at this age did not seem to understand the number condition of nonce words inflected with voiced plural /z/, or indeed even CVC singular nonce words.

These findings raise questions about how children acquire English plural morphology, and also raise questions about the role that allomorphic variation plays. At 24 months, children do have some understanding of plural morphology (which we should expect from studies of spontaneous speech [1], and also wug tasks [6]), but this understanding appears to be limited to the voiceless plural /s/. This does not look to be driven by language input, as plural allomorph /z/ accounts for over 70% of plurals children hear [1]. Potentially this may be driven by acoustic salience, as, in American English /s/ has been shown to have a longer frication duration than /z/ [10]. However, at this age, children do not appear to understand the number condition of novel CVC singular words. Perhaps children need to acquire an understanding of the complete set of English plural allomorphic variants before they are able to comprehend that the absence of a plural morpheme signifies singular. This is potentially demonstrated by the results of Experiment Two.

In Experiment Two, 36-month-olds were able to demonstrate an understanding of not only the syllabic plural /əz/, but an understanding of the singular as well. These results build upon the findings of previous IPL research [9], showing that 36-month-olds can, and do, comprehend all three English nominal plural allomorphs: /s/, /z/ and /əz/, despite the latter's relatively low frequency and proven difficulty in production tasks [11]. Furthermore, 36-month-olds' understanding of fricative-final singular nonce words suggests that they interpret plural morphology as *root+plural morpheme*, and are not simply parsing fricative-final words as being plural at this age.

These findings contribute to the small, but growing literature showing allomorphic variation affects children's acquisition of grammatical morphemes. A better understanding of this issue will help inform the gradual and variable nature of children's emerging linguistic abilities. It also provides a much-needed baseline against which to evaluate plural development in children developing bilingually, as well as with various types of language delay.

5. References

[1] Brown, R. (1973). A first language: The early stages. Cambridge, MA: Harvard University Press.

[2] De Villiers, J. G., & De Villiers, P. A. (1973). A cross-sectional study of the acquisition of grammatical morphemes in child speech. *Journal of Psycholinguistic research*, *2*(3), 267-278.

[3] Mervis, C. B., & Johnson, K. E. (1991). Acquisition of the plural morpheme: A case study. *Developmental psychology*, *27*(2), 222-235.

[4] Barner, D., Thalwitz, D., Wood, J., Yang, S. J., & Carey, S. (2007). On the relation between the acquisition of singular–plural morpho‐syntax and the conceptual distinction between one and more than one. *Developmental Science*, *10*(3), 365-373.

[5] Li, P., Ogura, T., Barner, D., Yang, S. J., & Carey, S. (2009). Does the conceptual distinction between singular and plural sets depend on language?. *Developmental psychology*, *45*(6), 1644.

[6] Berko, J. (1958). The Child's learning of English morphology. *Word, 14,* 150-177.

[7] Zapf, J. A., & Smith, L. B. (2007). When do children generalize the plural to novel nouns?. *First Language*, *27*(1), 53-73.

[8] Golinkoff, R. M., Hirsh-Pasek, K., Cauley, K. M., & Gordon, L. (1987). The eyes have it: Lexical and syntactic comprehension in a new paradigm. *Journal of child language*, *14*(1), 23-45.

[9] Kouider, S., Halberda, J., Wood, J., & Carey, S. (2006). Acquisition of English number marking: The singular-plural distinction. *Language Learning and Development*, *2*(1), 1-25.

[10] Smith, C. L. (1997). The devoicing of /z/ in American English: effects of local and prosodic context. *Journal of Phonetics*, *25*(4), 471-500.

[11] Mealings, K. T., Cox, F., & Demuth, K. (2013). Acoustic investigations into the later acquisition of syllabic -es plurals. *Journal of Speech, Language, and Hearing Research*, *56*(4), 1260-1271.

Exploring Articulatory Characteristics of Linking /ɹ/ in British English

Mitsuhiro Nakamura

Nihon University, Tokyo, Japan

nakamura.mitsuhiro99@nihon-u.ac.jp

Abstract

This study reports on the continuing investigation into the articulatory-acoustic nature of the linking /ɹ/ in British English. The linking /ɹ/ gesture was compared with the word-initial /ɹ/ gesture, using the EMA data collected from MOCHA-TIMIT. The results show that the linking /ɹ/s involve a smaller degree of tongue displacement and shorter articulatory durations than the word-initial /ɹ/s. This variation is closely related to the tongue configuration. The tip-up gesture of the linking /ɹ/ limits coarticulatory directionality to the horizontal dimension but the tip-down gesture to the vertical dimension. The results are discussed in terms of prosodic-positions and articulatory tasks.

Index Terms: Linking /ɹ/, EMA, coarticulation

1. Introduction

This exploratory study focuses on the articulatory characteristics of the linking /ɹ/ across word boundaries in British English. In standard British English and non-rhotic varieties of accent, a postvocalic /ɹ/ is not pronounced when followed by a word beginning with a consonant (e.g. *car park* [kɑː pɑːk]) but is pronounced when followed by a word beginning with a vowel (e.g. *car engine* [kɑːɹ endʒɪn]). This phenomenon is called linking /ɹ/. Linking /ɹ/ constitutes a part of the r-sandhi or r-liaison phenomenon which includes the intrusive /ɹ/ (e.g. *idea* [ɹ] *of*) and has been investigated from the point of view of phonetics and phonology (e.g. [1], [2], [3], [4]), sociolinguistics (e.g. [5]), and hiatus resolution (e.g. [6], [7]).

Compared to what we know about the word-initial /r/ (e.g. *research*), we have little knowledge about the articulatory characteristics of the linking /ɹ/. It is generally recognized that the word-initial /ɹ/ involves a primary constriction in the oral cavity and a secondary constriction in the pharyngeal cavity (e.g. [8]). Also, the tongue configuration varies from 'bunched/tip-down' to 'retroflexed/tip-up' (e.g. [9]). Are these features equally observed for the linking /ɹ/? What kind of features (if any) would differentiate the linking /ɹ/ gesture from the word-initial /ɹ/ gesture? These questions are explored in this paper. The production characteristics of the linking /ɹ/ are discussed in the light of prosodic positions [10] and constraints on 'articulatory tasks [11]'.

2. Method

2.1. Data collection and coding

The MOCHA-TIMIT database [12], which comprises articulatory (EPG, EMA, laryngograph (Lx)) and acoustic data of 460 sentences read by native speakers of English, was used for data collection. This study investigates the selected utterances spoken by three speakers of Southern British English referred to as SE (female), SA (male), and AP (male).

Collecting the speech materials took two steps. First, the potential linking /ɹ/ contexts were identified by an automatic search of relevant spellings appearing in the MOCHA 460 sentences (e.g. 'er' in *higher*, 'ur' in *your*). The cases followed by a space and a vowel letter were extracted. This process produced 70 potential contexts of a linking /ɹ/ across word boundaries (e.g. *numbe[ɹ] of, fo[ɹ] hours, you[ɹ] interview*). Similarly, searching the word-initial /r/s preceded by a space and a vowel letter yielded 18 tokens (e.g. *The rich, occasionally reads, to wrap*). Secondly, those 70 potential contexts of a linking /r/ were analyzed using auditory and acoustic methods. All the cases identified as containing a linking /ɹ/ were confirmed by examining a low F3 or trough on the spectrogram. The number of the linking /ɹ/ realizations varies with the speakers. 12 instances (17%) were found for SE, 51 instances (72%) for SA, and 32 instances (45%) for AP.

In order to examine the effects of linguistic features, each vowel flanking the linking /r/ was coded by the backness (front, central, back) and height (close, mid, open).

2.2. Measurements

The measurements of the articulatory data were conducted where tokens of word-initial /ɹ/ and linking /ɹ/ were realized.

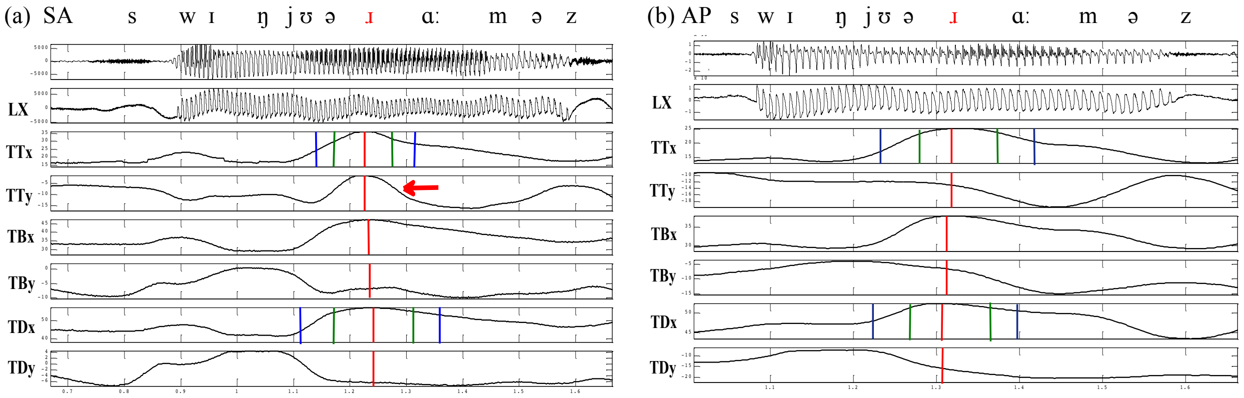


Figure 1: *Measurement points for TT, TB, and TD trajectory*

Figure 1(a,b) are representative examples of linking /ɹ/ in the sentence ‘Swing you[ɹ] arms as high as you can.’ Figure 1 also illustrates the measurement points used.

For the analysis of spatial characteristics, the horizontal (x) and vertical (y) displacements were recorded for the tongue tip (TT), tongue body/middle (TB), and tongue dorsum (TD) at the tangential velocity minimum point (the red line in Figure 1). The TT coil was placed 7-10mm back from the extended tongue tip, the TB coil was placed 20-30mm back from the TT coil, and the TD coil was placed 20-30mm back from the TB coil.

Temporal characteristics were analyzed for the movement of the tongue tip and the tongue dorsum. The duration measurement in this study was based on the four time points (onset, target, release, release offset) proposed by [13]. These points were specified at 20% ([14]) of the tangential velocity peak associated with movement towards, or away from, a target constriction. The two articulatory durations were identified. The *total* duration is the time interval between the onset and the release offset of a given movement (an interval marked by blue lines in Figure 1). The *plateau* duration is the interval between the target and the release of a given movement (an interval marked by green lines in Figure 1).

Statistical comparisons (t-test and one-way ANOVAS) were performed, using SPSS©, separately for the two speakers SA and AP. SE was excluded since the number of linking /r/ tokens was small.

3. Results

3.1. Linking /ɹ/ vs. word-initial /ɹ/

3.1.1. Identifying the gestures

Before presenting the results, it would be worthwhile to point out some characteristics captured in Figure 1(a,b). At the velocity minimum point (marked by a red line), all the three coils, TT, TB, TD, reveal a certain amount of retraction (i.e. the distinctive movement in the x dimension). However, the raising of the TT (i.e. the distinct movement in the y dimension, TTy, indicated by an red arrow in Figure 1(a)) is found in SA’s production only. This feature, which typifies individual differences, will be clarified below.

3.1.2. Spatial aspects

Figure 2 presents all the measurements of the horizontal (x) and vertical (y) displacement of TT, TB, TD for the linking /ɹ/ and word-initial /ɹ/ for the three speakers. The general trend is that the plot pattern for the linking /ɹ/ is not substantially divergent from that of the word-initial /ɹ/ within a speaker. However, clear patterns emerge between the speakers. The TT position is kept higher in SA’s productions but is lower in AP’s production. It is difficult to specify the exact tongue shapes based on the EMA data but the differences captured in Figure 2 can be labeled as the tip-up gesture for SA and as the tip-down gesture for AP.

Table 1: *Mean displacement of linking and initial /ɹ/* (mm)

		SA (s.d.)	AP (s.d.)
TTx	Linking r	**28.7** (2.3)	28.1 (1.3)
	Initial r	**32.8** (4.2)	26.5 (3.1)
TTy	Linking r	-1.1 (1.9)	-8.6 (1.9)
	Initial r	-0.2 (0.8)	-10.0 (1.7)
TBx	Linking r	**41.6** (1.9)	40.3 (0.9)
	Initial r	**44.5** (2.9)	39.4 (2.1)
TBy	Linking r	-2.0 (1.5)	-7.7 (1.5)
	Initial r	-3.7 (1.9)	-9.3 (3.0)
TDx	Linking r	**53.7** (2.1)	53.9 (1.2)
	Initial r	**56.5** (1.8)	52.7 (2.0)
TDy	Linking r	**-2.0** (1.7)	**-14.7** (1.2)
	Initial r	**-4.6** (1.9)	**-17.5** (3.2)

*For SA, initial r, n=4; linking r, n=12, and for AP, initial r, n=4; linking r, n=8.

To examine whether the tongue positions of the linking /ɹ/ are different from those of the word-initial /ɹ/, the tokens whose flanking vowels are schwas (including a diphthong [əʊ]) were selected: e.g., for the linking /r/, *December and*, *are open*, *Regular attendance*; for the word-initial /r/, *a romantic*, *the rose*. The mean displacement of the TT, TB, TD is given in Table 1. Each position of the tongue was separately examined by t-test (2 tailed). Significant differences are marked red in Table 1.

For SA, the tongue tip and body is significantly more retracted (but not raised) for a word-initial /ɹ/ than for a linking /ɹ/, and the tongue dorsum is more retracted and raised for a word-initial /ɹ/: TTx [t=-2.495, df=14, p=0.026], TBx [t=-2.326, df=14, p=0.036], TDx [t=-2.474, df =14, p=0.027], and TDy [t=2.552, df=14, p=0.023]. For AP, the tongue dorsum is more raised for a word-initial /ɹ/: TDy [t=2.268, df=10, p=0.047].

These results suggest that there is a difference between a linking /ɹ/ and a word-initial /ɹ/ on the one hand and a difference between the tip-up and the tip-down configuration of the tongue on the other.

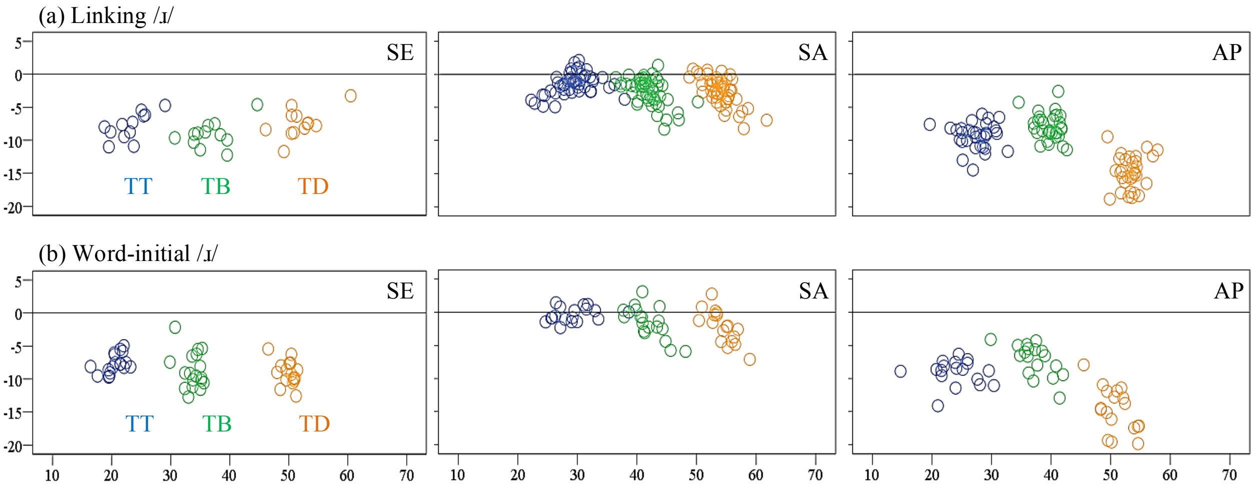


Figure 2: *TT, TB, and TD displacement of linking /ɹ/ and word-initial /ɹ/ for the three speakers* (mm)

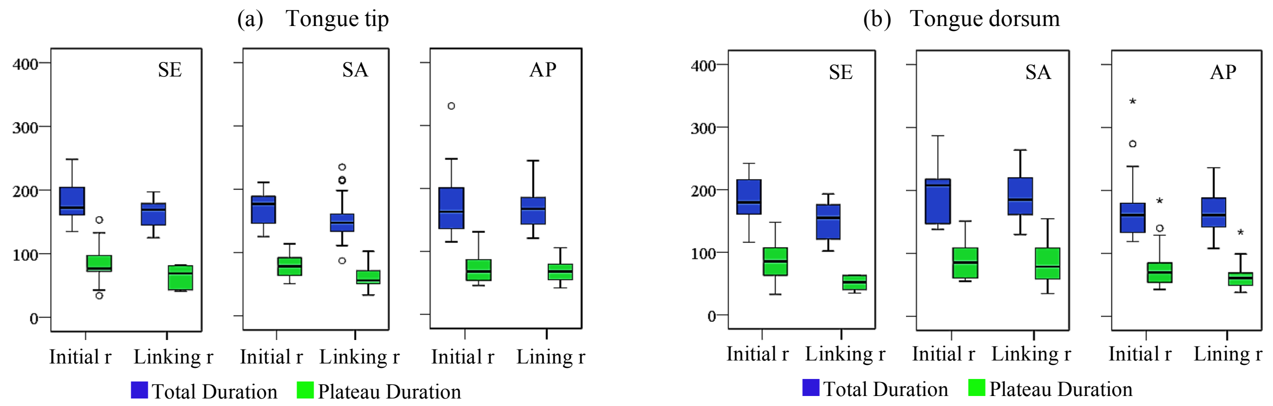

(a) Tongue tip (b) Tongue dorsum

Figure 3: *Mean total and articulatory plateau duration of linking /ɹ/ and word-initial /ɹ/*

3.1.3 *Temporal aspects*

We now move on to the temporal aspects of the linking /ɹ/ and the word-initial /ɹ/. Figure 3 (a,b) summarizes the mean total and articulatory plateau durations of the tongue tip and the tongue dorsum for the three speakers. The three speakers show a general trend that the two durations of the linking /ɹ/ are shorter than those of the word-initial /ɹ/. Comparisons were made separately for SA (linking /ɹ/, n=39; word-initial /ɹ/, n=18) and AP (linking /ɹ/, n=26; word-initial /ɹ/, n=18) by t-test (2 tailed).

SA shows a significant difference in the durations of the tongue tip gesture (TT total [t=-2.577, df=55, p=0.013] and TT plateau [t=-3.714, df=55, p<0.0001]), but not in those of the tongue dorsum gesture (TD total [t=-0.456, df=55, p=0.650] and TD plateau [t=-0.595, df=55, p=0.554]). Thus, the duration of the tongue tip retracting (and raising) gesture is shorter in the linking /ɹ/ but not that of the tongue dorsum gesture.

In contrast, AP reveals no significant differences: for the tongue tip, TT total [t=-0.567, df=42, p=0.574] and TT plateau [t=-0.676, df=42, p=0.503]; and for the tongue dorsum, TD total [t=-0.875, df=42, p=0.387] and TD plateau [t=-1.702, df=42, p=0.96]. It might be possible to assume that this result is related to the configuration of the /ɹ/ gesture used by speaker AP. However, there is still too much uncertainty to confirm such an interpretation. Further research is necessary on this point.

Table 2: *Mean displacement and effects of V2 backness* (mm)

(a) SA

	Front (s.d.)	Central (s.d.)	Back (s.d.)	F(2,42)
TTx	28.3 (2.8)	29.0 (2.3)	31.2 (4.2)	3.05
TTy	-1.6 (1.6)	-1.0 (1.8)	-1.8 (1.1)	0.83
TBx	41.0 (2.0)	41.8 (2.0)	43.5 (3.7)	**3.35**
TBy	-2.1 (1.2)	-2.2 (2.0)	-3.7 (2.7)	2.61
TDx	53.4 (1.9)	53.6 (2.0)	55.7 (3.4)	**3.71**
TDy	-2.0 (1.4)	-2.3 (2.0)	-3.6 (3.2)	2.13

(b) AP

	Front (s.d.)	Central (s.d.)	Back (s.d.)	F(2,25)
TTx	26.9 (3.7)	28.0 (1.6)	27.1 (2.6)	0.47
TTy	-9.0 (1.4)	-9.3 (2.3)	-10.1 (2.2)	0.55
TBx	39.0 (2.2)	40.2 (1.1)	39.5 (1.9)	1.50
TBy	-7.8 (2.1)	-7.7 (2.4)	-7.8 (1.0)	0.01
TDx	52.6 (2.0)	53.7 (1.7)	53.0 (1.2)	1.08
TDy	-14.1 (2.7)	-14.9 (2.3)	-15.6 (1.3)	0.82

*For SA, front, n=22; central, n=14; back, n=9; For AP, front, n=10; central, n=12; back, n=6

3.2. Effects of V2 backness and height on the TT, TB, TD displacements of linking /ɹ/

In order to further explore the articulatory nature of the tip-up (SA) and the tip-down (AP) gesture, the effects of the backness and height of the following vowels (V2) was examined. In this analysis, the vowel preceding the linking /ɹ/ (V1) is fixed to a schwa (including [aʊə] and [eə]): e.g., *for eating, sculpture in, ever enter, power outage, their own, thermometer under, paper and, Her auburn, oyster on*. We will examine how the two kinds of linking /ɹ/ gesture accommodate to coarticulatory effects of the changing V2s.

Table 2 summarizes the mean displacements in terms of the backness categories and Table 3 in terms of the height categories. The results of one-way ANOVAS are also presented in the right-most column of each table and significant differences are marked red.

In the productions by SA, as shown in Table 2(a), the V2 backness effects were found to be significant in the horizontal dimension of the tongue body and dorsum (TBx, p=0.045; TDx, p=0.033). The TB and TD of the linking /ɹ/ are more retracted when followed by a back vowel than when followed by a front or central vowel. Note in passing that the horizontal movement of the tongue tip (TTx) is virtually significant (p=0.058). In contrast, as summarized in Table 3(a), no significant

Table 3: *Mean displacement and effects of V2 height* (mm)

(a) SA

	Close (s.d.)	Mid (s.d.)	Open (s.d.)	F(2,42)
TTx	27.7 (3.4)	29.2 (2.4)	30.2 (3.3)	2.62
TTy	-1.8 (1.8)	-1.0 (1.7)	-1.7 (1.2)	1.30
TBx	40.6 (2.3)	41.8 (1.9)	42.6 (2.9)	2.54
TBy	-1.9 (1.1)	-2.2 (1.5)	-3.2 (2.5)	2.10
TDx	52.9 (2.2)	53.9 (2.2)	54.7 (2.6)	2.05
TDy	-1.8 (1.3)	-2.1 (1.8)	-3.2 (2.7)	1.85

(b) AP

	Close (s.d.)	Mid (s.d.)	Open (s.d.)	F(2,25)
TTx	25.7 (4.3)	28.1 (1.8)	27.6 (2.1)	1.77
TTy	-8.6 (1.6)	-9.0 (1.8)	-10.1 (2.2)	1.64
TBx	38.2 (2.6)	40.1 (1.2)	39.8 (1.4)	2.89
TBy	-7.3 (2.4)	-7.8 (1.4)	-8.0 (2.4)	0.23
TDx	52.4 (2.6)	53.7 (1.2)	53.1 (1.7)	1.04
TDy	-12.4 (1.7)	-15.0 (1.4)	-15.8 (2.5)	**6.00**

*For SA, close, n=13; mid, n=16; open, n=16; For AP, close, n=6; mid, n=11; open, n=16 differences were found for the effects of the V2 height.

Conversely, AP indicates a significant difference in the effects of the V2 height but not in those of the V2 backness. As seen in Table 3(b), the significant difference is limited only to the vertical dimension of the tongue dorsum (TDy, p=0.007). The dorsum is significantly higher when followed by a close vowel than when followed by a mid or open vowel.

4. Discussion

We have so far investigated the spatiotemporal characteristics of the /ɹ/ gesture in the production of a linking /ɹ/ and a word-initial /ɹ/. The results presented above are summarized as follows: (i) the constrictive approximation for a linking /ɹ/ is similar to that for a word-initial /ɹ/. One speaker (SA) tends to use the tip-up configuration and the other speaker (AP) the tip-down configuration. (ii) The TT, TB, TD displacements of a linking /ɹ/ are smaller in degree than those for a word-initial /ɹ/. This variation is restricted to the horizontal dimension in the tip-up articulation and to the vertical dimension in the tip-down articulation. (iii) In the tip-up production, the total and articulatory plateau duration of a linking /ɹ/ is shorter in the TT gesture (but not in the TD gesture) than that of a word-initial /ɹ/, while such durational variations are not significant in the tip-down production. And (iv) for the coarticulatory effects of the following vowel (V2) on the linking /ɹ/ gesture, the effects of the V2 backness (but not height) is significant in the tip-up production (SA) and the TB and TD gesture is more retracted when followed by a back vowel than when followed by a front or central vowel. In contrast, in the tip-down production (AP), the effects of the V2 height (but not backness) is significant and only the TD gesture is more raised when followed by a close vowel than when followed by a mid or open vowel.

Here we shall focus on the spatial differences between the linking /ɹ/ and the word-initial /ɹ/. This can be interpreted as the effects of the position-related, or the domain initial, articulatory strengthening [10]. The spatial positioning becomes more extreme in the production of word-initial /ɹ/s. However, it is evident in the results given in Table 1 that such a strengthening is not applied equally to the two dimensions of a given gesture. The tip-up type constriction gesture allows the strengthening in the horizontal (x) dimension of the tongue but not in the vertical (y) dimension. In contrast, the tip-down constriction gesture accepts the strengthening of the tongue dorsum in both dimension, but not that of the tongue tip and body. Similar results are obtained for the effects of V2 backness and height in the production of linking /ɹ/s (Tables 2 & 3). The results reflect the distinctive properties of the tip-up and tip-down gestures, as well as the position-related spatial variation and the V2 backness/height-related effects.

Why do the tip-up and the tip-down constrictions differ in their realization of the position-related effects and the V2 coarticulatory effects? Given that the F3 lowering is a distinct characteristic of /ɹ/ and is a front cavity resonance [15], both constriction gestures aim at making a space in the front of the oral cavity. Coarticulatory directionality limited to a single dimension of the articulatory movement, therefore, could be interpreted as one realizational strategy specific to the given constriction. This assumption will be substantiated by further research of an articulatory-acoustic analysis of the linking /r/ and the word-initial /r/.

Finally, some comments should be made regarding speaker SE's phonetic realizations in the potential linking /ɹ/ environment. In the process of extracting the linking /ɹ/ tokens, variable realizations were coded by the following four categories: (i) linking /ɹ/, (ii) hiatus, (iii) glottalized, and (iv) other. The glottalized realizations, which are subcategorized into 'glottal stop and creaky voice,' are the commonest in SE's production (60%, 42/70 tokens). Also, AP shows 'other' patterns (e.g. a weak vowel/syllable deletion) more frequently than the other speakers (29%, 20/70 tokens). The issue of hiatus resolution, which involves speakers' strategies to create a smooth transition, will be investigated in future research.

5. Conclusions

The current study has explored the articulatory nature of the linking /ɹ/ productions across word boundaries in British English. This study has used the data collected from the multichannel articulatory database. Controlled experiments are also necessary to substantiate the articulatory patterns we have explored. It is hoped, however, that the parametric phonetic analysis presented in this study provides another interesting description of the linking /ɹ/.

6. Acknowledgements

I am grateful to Valerie Hazan and Mark Huckvale of University College London, and two anonymous reviewers of SST2016 for their criticisms, suggestions, and encouragement. All errors are mine. This research is supported in part by a grant from Japan Society for the Promotion of Science, Grant-in-Aid for Scientific Research (C) 2016, Grant Number 15K02491.

7. References

[1] Mompean, J. and Mompean, P., "/r/-liaison in English: An empirical study", Cog. Ling., 20: 733-776, 2009.

[2] Mullooly, R., "An Electromagnetic Articulograph study of alternating [r] and the effects of stress on rhotic consonants", PhD dissertation, Queen Margaret University College, 2004.

[3] McMahon, A., Foulkes, P. and Tollfree, L., "Gestural Representation and Lexical Phonology", Phonol, 11(2): 277-316, 1994.

[4] Gick, B. "A gesture-based account of intrusive consonants in English", Phonol, 16(1): 29-54, 1999.

[5] Foulkes, P. "English [r]-sandhi: a sociolinguistic perspective", Histoire Épistémologie Langage, 19(1): 73-96, 1997.

[6] Britain, D. and Fox, S., "The Regularisation of the Hiatus Resolution System in British English", in M. Filppula et al. [Ed], Vernacular Universals and Language Contacts: Evidence from Varieties of English and Beyond, 177-205, Routledge, 2009

[7] Cox, F., Palethorpe, S., Buckley, L. and Bentink, S., "Hiatus resolution and linking 'r' in Australian English", JIPA, 44(2): 155-178, 2014.

[8] Alwan, A., Narayanan, S., and Haker, K. "Toward articulatory-acoustic models for liquid approximants based on MRI and EPG data. Part II. The rhotics", JASA, 101(2): 1078-1089, 1997.

[9] Delattre, P. and Freeman, D.C. "A Dialect Study of American R's by X-ray Motion Picture", Linguistics, 6(44): 29-68, 1968.

[10] Keating, P., Cho, T., Fougeron, C., and Hsu, C.S. "Domain-initial articulatory strengthening in four languages", in J. Local, R. Ogden, and R. Temple [Eds], Papers in Laboratory Phonology VI: Phonetic Interpretation, 145-163, Cambridge Univ. Press, 2004.

[11] Browman, C. and Goldstein, L. "Articulatory Phonology: An Overview", Phonetica, 49: 155-180, 1992.

[12] Wrench, A., "MOCHA-TIMIT", The Centre for Speech Technology Research, University of Edinburgh. Online: http://www.cstr.ed.ac.uk/research/projects/mocha.html, accessed on 21 Jan 2015.

[13] Gafos, A.I. "A Grammar of Gestural Coordination", NLLT, 20: 269-337, 2002.

[14] Hoole, P. "Issues in the acquisition, processing, reduction and parameterization of articulatory data", FIPKM, 34: 158-173, 1996.

[15] Espy-Wilson, C.Y., Boyce, S.E., Jackson, M., Narayanan, S., and Alwan, A. "Acoustic modeling of American English /r/", JASA, 108(1): 344-356, 2000.

Tongue positions corresponding to formant values in Australian English vowels

Arwen Blackwood Ximenes[1], *Jason A. Shaw*[1], *Christopher Carignan*[1]

[1]The MARCS Institute, Western Sydney University

a.blackwoodximenes@westernsydney.edu.au

Abstract

A common assumption about vowel formants is that F1 inversely correlates with tongue height and F2 inversely correlates with tongue backness. This study compared vowel formants and corresponding lingual articulation in Australian English (AusE) for nearly all of the AusE monophthongs. Simultaneous acoustic and electromagnetic articulography (EMA) recordings are reported for four speakers producing multiple repetitions of ten monophthongs. Results show that, while in general formants correspond to the articulatory data, there are also cases in which the typically assumed correspondence breaks down. Consistency in Tongue Dorsum position was observed despite variation in F2.

Index Terms: speech production, acoustics, Electromagnetic articulography, Australian English

1. Introduction

Assumptions are commonly made regarding the articulatory nature of vowels based on their formant values: F1 is assumed to be inversely correlated with tongue height and F2 is assumed to be inversely correlated with tongue backness. This paper assesses this correspondence, reporting the tongue position of Australian English (AusE) vowels and corresponding formant values.

There is an abundance of acoustic studies of AusE vowels but comparative articulatory data are lacking. Some recent studies on vowel articulation focus on a small subset of vowels. Tabain [1] investigates the articulatory and acoustic properties of one vowel in different prosodic contexts. Watson, Harrington and Palethorpe [2] compared the acoustic and articulatory vowel spaces of AusE and New Zealand English (NZE). Their analysis covered four vowels, those in the words *hid*, *head*, *had*, and *herd*. Lin, Palethorpe and Cox [3] looked at a larger number of AusE vowels in the /CVl/ context. They focused on how vowel height influences lateral production (/CVl/) rather than on the phonetic properties of the vowels themselves. The degree to which the following /l/ influences the preceding vowel is not clear. The most comprehensive articulatory study of AusE vowels was undertaken over four decades ago [4]. Bernard reports on the results of an X-ray study investigating all the AusE vowels but does not report any quantitative measurements of the data. Bernard's qualitative description of X-ray data still constitutes the most comprehensive analysis of Australian vowel articulation to date.

This paper aims to address the lack of quantitative data on AusE vowels by describing the AusE vowel space articulatorily and comparing the results to formant values.

2. Method

2.1. Subjects

Articulatory and acoustic data were analysed from four Australian English speakers (two males and two females) ranging in age at time of recording from 20 to 42. All participants were recruited from the Western Sydney University community.

2.2. Materials

Stimuli comprised a list of lexical items and nonce words containing 15 vowels, including 10 monophthongs, in the sVd context. This paper focuses on analysis of the monophthongs. We list the stimulus items below. Each item is followed by, in parentheses, the reference word for the vowel devised by Wells [5]. The reference word disambiguates the spelling, which is particularly useful for nonce words: *said* (DRESS), *seed* (FLEECE), *sood* (FOOT), *sued* (GOOSE), *sid* (KIT), *sod* (LOT), *sawed* (THOUGHT), *surd* (NURSE), *sud* (STRUT), *sad* (TRAP). This set of monophthongs covers the whole AusE acoustic vowel space. The only AusE monophthong missing is START, which according to Cox [6] does not differ in its formants from STRUT.

2.3. Procedure

The movements of the articulators were tracked using an Northern Digital Inc. Wave EMA system at a sampling rate of 100Hz. This system uses an electromagnetic field to track the movement of small receiver coils or sensors (3 mm in size) glued or taped to the articulators. The electromagnetic field induces an alternating current in the sensors, and the strength of this current is used to determine the position of the sensors in relation to the transmitter. Articulatory movements are captured in the vertical, horizontal and lateral dimensions with high spatial-temporal resolution (< 0.5 mm rms error). In this study, we focused on movements in the horizontal and vertical dimension, since these are the dimensions typically assumed to correspond to formant values. The sensor trajectories were synchronized to the audio signal during recording by the NDI system. EMA sensors were glued to the following articulators along the midsagittal plane: jaw (below the lower left incisor), lips (at the vermillion edge of the upper and lower lip), tongue tip (TT), tongue blade (TB) and tongue dorsum (TD). The TD sensor was placed as far back as comfortable for the participant. The TT sensor was placed near the tip and the TB sensor was placed midway between the TT and TD sensors.

The target stimulus words were displayed on a computer monitor placed outside of the magnetic field. One word was presented per trial. There were 15 trials (one per vowel) per block and eight blocks in the experiment. This resulted in 15 (vowels) x 8 (repetitions) = 120 vowel tokens per participant.

Of the recorded data, the monophthongs consist of 10 (vowels) x 8 (repetitions) = 80 tokens per participant, 320 monophthong tokens in total. Speech acoustics were recorded using a shotgun microphone at a sampling rate of 22 kHz. Technical problems due to data acquisition, analysis, and mispronunciation, resulted in three tokens (less than 1% of the total data) being excluded from the analysis.

Head movements were corrected computationally, with reference to sensors glued to the nasion and mastoids. The articulatory data were rotated relative to the occlusal plane so that the origin of the coordinate system corresponds to the front teeth. The occlusal plane was established by having the participant bite down on a protractor with 3 sensors affixed in a triangular formation.

2.4. Articulatory measurements

Measurements were extracted from sensor trajectories using the labelling procedure, *findgest*, an algorithm developed for the Matlab-based software package, “Multi-channel visualization application for displaying dynamic sensor movement” (MVIEW), by Marke Tiede at Haskins Laboratories. This program was used to detect the nearest tangential velocity minimum of the TD sensor (taken from movement in the horizontal and vertical dimensions together) during the interval corresponding to the vowel. We then extracted positional coordinates from all the lingual sensors at this vowel target landmark.

In some cases the parse MVIEW provided for the TD velocity minimum was impacted on by the surrounding consonants, i.e. TD reached its positional target for the vowel during the preceding /s/ (velocity minimum), making it difficult to differentiate the velocity peaks of the vowel and adjacent consonant. TB was used in cases where it showed more controlled movement towards vowel constriction.

2.5. Acoustic measurements

Formant data (F1 and F2) for each vowel was extracted from the sound files at the time point of the articulatory measurements (i.e., the vowel target, as described above). Using the time points extracted from the articulatory measures for the acoustic analysis enables a direct comparison between articulation and acoustics. Our method of parsing vowel targets using the point of minimum velocity in the articulatory data follows similar general principles used to identify formants in Cox [6] and Harrington, Cox and Evans [7]. In these papers vowel targets were identified based on formant displacement patterns, e.g., max/min F1/F2, depending on vowel. Max/min formant values correlate closely to the minimum velocity of articulator movement in our data. Other acoustic studies have used the acoustic midpoint of the vowel, which did not consistently correspond to the velocity minimum of the TD or TB sensors in this data.

2.6. Analysis

One of the challenges of analyzing speech production across speakers is that anatomical differences influence both the formant values and EMA positional coordinates. In the case of formants, differences in vocal tract length influence the average formant values. In articulatory data, differences in tongue shape, volume, and sensor placement lead to different average values. In both cases, because of differences in anatomy, between-speaker differences for the same vowel can be larger than within-speaker differences across vowels. In order to facilitate comparison across our four speakers, we normalized both the formant values and the positional coordinates by calculating z-scores of the horizontal and vertical dimensions for the TD sensor and of F1 and F2. Z-scoring preserves the within-speaker structure of the data but allows for a direct comparison across speakers by controlling for interspeaker vocal tract differences.

3. Results

We report the acoustic results first followed by the articulatory results. After describing the general pattern in both sets of data and the correspondence between them, we take up some exceptions to the main pattern in the discussion section.

3.1. Acoustic data

The distribution of normalized formant values (F1 and F2) across the acoustic vowel space for all speakers is presented in Figure 1. The ellipses contain 95% of the data for each vowel, and are centered on the mean of each vowel category.

In line with previous acoustic studies of AusE (e.g., [6]), the vowels are fairly evenly distributed across the vowel space and can be classified as “front”, “central”, and “back” on the basis of the formants. There are four vowels with high F2 (i.e. “front” vowels) that differ in F1: FLEECE, KIT, DRESS and TRAP. There are also differences in F2 amongst the front vowels, but part of these differences are due to general properties of formant spaces, e.g., as F1 increases, F2 of front vowels decreases. We assume that the differences in F2 are at least in part attributable to this relationship and may not be under speaker control. There are three “central” vowels that have intermediate F2 values, GOOSE, NURSE and STRUT, and also differ in increasing F1. The remaining “back” vowels have low F2: FOOT, LOT and THOUGHT. We now turn to the articulatory data to observe how the differences in formant values correspond to tongue position.

3.2. Articulatory data

In order to assess whether the observations of formants correspond to “front”, “central”, and “back” articulatory positions, we focus on the TD sensor. The mapping from articulation to acoustics is of course impacted by differences in vocal tract diameter across the entire length of the vocal tract. Nevertheless, we have found that even data from the TD sensor alone reveals a general correspondence to formant values in line with expectations. Figure 2 shows the normalized values (z-scores) of the TD sensor for all four subjects. The TD data represents the range of motion with which that fleshpoint on the tongue varies across vowels. The *y*-axis shows the vertical position, and the *x*-axis shows horizontal position from front (positive z-scores on the left side of the figure) to back (negative z-scores on the right side of the figure). As with the formant data, ellipses contain 95% confidence intervals for each vowel distribution and are centered on the mean.

The distribution of vowels in the articulatory data generally follows the distribution of vowels in formant space. More specifically, F1 tends to be inversely correlated with tongue height, and F2 tends to be inversely correlated with tongue backness. Of the “front”, “central”, and “back” vowels determined on the basis of the formants, the back vowels show the least overlap at the TD sensor. The “back” vowels, FOOT, THOUGHT, and LOT, have a TD position more posterior than the other vowels, as indicated by the negative z-score.

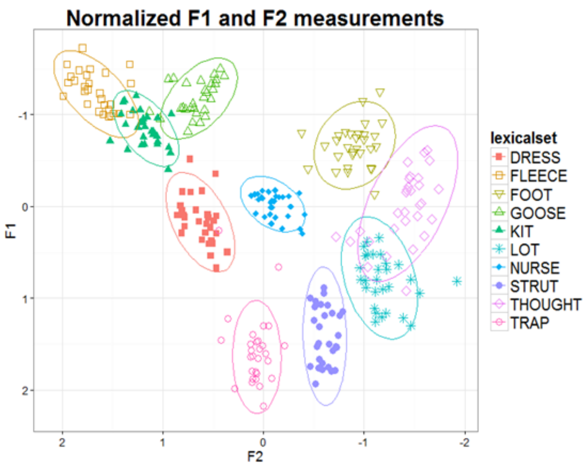


Figure 1. *Normalised F1 and F2 for Australian English vowels*

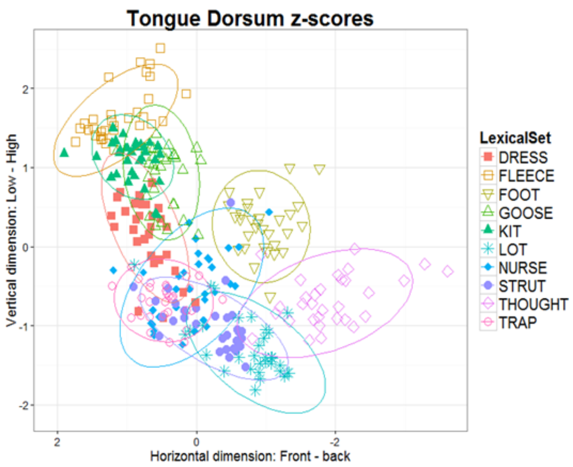


Figure 2. *Z-scores of the Tongue Dorsum sensor position for Australian English vowels*

The center of the ellipses for STRUT and NURSE are closest to zero on the *x*-axis, indicating that they are at the average level of backness in the data. These vowels, in addition to GOOSE, are the “central” vowels: they all had intermediate F2 values. Of these three central vowels, GOOSE is the most front TD position. Presumably, rounding of GOOSE lowers F2, compensating for TD frontness. The front vowels FLEECE, KIT, DRESS, and TRAP have horizontal positions that are higher than all the other vowels, indicating that they have the lowest degree of backness.

Although Figure 2 shows data from just a single fleshpoint on the tongue, articulatory differences that correspond to those in the formants can be observed. In particular, the relative height and backness of vowels at the TD sensor is preserved in the F1 and F2 values. The vowel space expressed in terms of TD position is more compact than the vowel space expressed in formants. Consequently, there is more overlap in the articulatory data compared to the acoustic data. From this we can ascertain that other aspects of vowel articulation function to enhance the differences observable from TD position, so TD does not capture all of the articulatory change.

4. Discussion

The AusE vowels in this study can be clearly differentiated on the basis of F1 and F2, and a similar partitioning of the vowel space can be observed in the position of the TD sensor in vertical and horizontal dimensions. The AusE vowel space can be viewed as taking a 4:3:3 configuration, whereby there are four “front” vowels differing in height, three “central” vowels differing in height and “three” back vowels also differing in height. For the most part, the acoustic and articulatory data are in correspondence, as is expected from the assumption that F1 is inversely correlated with tongue height and F2 is inversely correlated with backness. This is remarkable given that the articulatory data come from a single fleshpoint, the TD. It is important to note that this relationship suggested requires more precise quantification. Incorporating other aspects of articulation, in particular jaw height, and tongue curvature may provide a more dispersed view of the articulatory vowel space. The overlap seen at the TD for some vowels may be unimportant if those vowels are differentiated in another part of the vocal tract, such as is suggested by Wood [8], where four different constriction locations are proposed.

Alongside the general correspondence observed across Figure 1 and Figure 2 in relative position of vowels, we have also identified some mismatches. When we zoom in on individual speakers, we observe some cases in which a change in F2 does not correspond with differences in TD position or in the position of other lingual sensors, as expected

For one male speaker, we found a mismatch in backness for vowels LOT and THOUGHT. The general trend in the data is for the following correspondence: F2, lower for THOUGHT than for LOT, corresponds to TD position, which is also further back for THOUGHT than for LOT. One speaker shows the group pattern in TD position (TD further back for THOUGHT than for LOT) but does not show the group pattern in F2. Rather, for this speaker, the F2 for THOUGHT was not lower than for LOT (leading to some overlap between the THOUGHT and LOT ellipses in Figure 1). It is possible that this is due to reduced rounding in THOUGHT for this speaker. Although we have not as yet been able to quantify the effect, rounding is expected to be greater for THOUGHT than for LOT and should contribute to the separation in F2. Cases such as this underscore the indeterminacy of interpreting formant values in terms of articulation, or at least on a single fleshpoint. Because they are shaped by multiple articulatory constrictions in the vocal tract, it is not always possible to map changes in formants to changes in TD position. In this case, articulation shows consistency across speakers while formant values show variation, which is likely attributable to degree differences in rounding.

Another mismatch in the data is less easy to explain. In the other of our male speakers, we observed an inconsistency in the acoustic-articulatory relation in the “central” part of the vowel space. Although this speaker shows the same level of correspondence as other speakers in the front and back

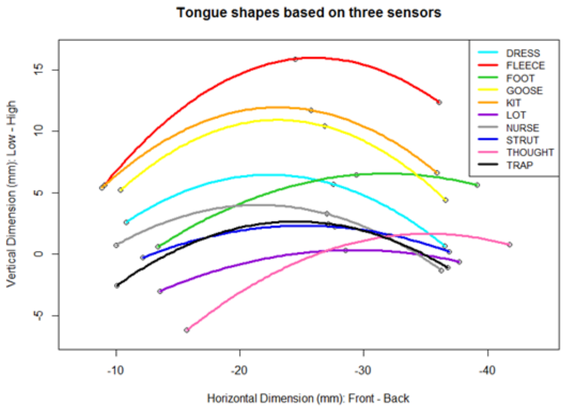


Figure 3. *Averages of a male speaker's three lingual sensors, with polynomials fit to the averaged sensor points for each vowel. The tongue tip is on the far left and the tongue dorsum is on the far right.*

sections of the vowel space, the central vowels NURSE, GOOSE and STRUT all have a similar level of backness (i.e., horizontal position of the TD), but there are large differences in F2. Despite similar horizontal positions of the TD (and TB and TT), F2 is highest for GOOSE, followed by NURSE, then STRUT. The averages of this speaker's lingual sensors are represented in Figure 3, with a polynomial fit to the three sensor points for each vowel.

Unlike the case of THOUGHT and LOT discussed above, it is unlikely that the difference in F2 across GOOSE, NURSE and STRUT is due to a degree difference in rounding. Lip rounding is expected to be the greatest degree for GOOSE, followed by NURSE. Bernard [4] reported smaller lip aperture for GOOSE than NURSE). However for this speaker, GOOSE and NURSE show unexpectedly high F2 values. Rounding would be expected to lower F2, the opposite pattern of what we observed. One hypothesis for this mismatch between acoustic and articulatory data is that the relation between F2 and backness is nonlinear in this portion of the vowel space, i.e. sometimes small differences in backness may have a large influence on F2 [9]. Given the particular anatomy of this speaker, central vowels may have unstable relations between F2 and TD backness. An alternative hypothesis is that something else is influencing F2 other than TD backness. One possibility may be the differences in tongue curvature which can be seen in Figure 3, and which has been shown to differentiate the vowels of English [10]. Another suggestion is that height influences F2 to a greater degree than backness in the central vowel space for this speaker. A third possibility is that aspects of lingual articulation outside of the mid-sagittal plane, e.g., tongue grooving, may be playing a role. A fourth possibility is lip rounding. Further investigation would be needed to discover why F2 varies despite similar degrees of TD backness for these central vowels, and why this is the case in this part of the vowel space and for this speaker in particular.

5. Conclusions

Generally speaking, the relationship between acoustics and articulation as previously described, such that F1 is inversely related to vowel height and F2 is inversely related to backness, was confirmed in our report of Australian English monophthongs. Moreover, the relationship was apparent from a single fleshpoint on the tongue, attached to the Tongue Dorsum, although a more precise quantification of the relation will require incorporating other dimensions of articulation. There were also a few corners of the data in which the assumed correspondence between acoustics and articulation broke down. For one speaker, the backness of the central vowels did not correspond to F2. For another, the backness of back vowels did not correspond to F2. In both cases, we observed consistency in TD position across vowels despite variation in F2. In the latter case but not the former, it is likely that rounding perturbs the relation between TD backness and F2. We conclude that formant values offer a heuristic for diagnosing TD position on the basis of acoustic data which is largely valid, particularly for front vowels and where vowel rounding is not at issue. For some speakers, F2 may not provide a valid indication of TD backness for central vowels, although additional research is needed to understand the precise conditions under which the normally assumed correspondence between F2 and TD backness breaks down.

6. Acknowledgements

This research was supported in part by Australian Research Council grant DE120101289 to the second author. We would like to thank all the speakers for their participation and colleagues at the MARCS Institute for their technical assistance with EMA experiments.

7. References

[1] Tabain, M. (2008). Production of Australian English language-specific variability. *Australian Journal of Linguistics, 28*(2), 195-224.

[2] Watson, C. I., Harrington, J., & Palethorpe, S. (1998). A kinematic analysis of New Zealand and Australian English vowel spaces. In ICSLP.

[3] Lin, S., Palethorpe, S., & Cox, F. (2012). An ultrasound exploration of Australian English /CVl/ words. *Proceedings of the 14th Australasian International Conference on Speech Science and Technology*, p.105-108.

[4] Bernard, J. B. (1970). A cine-X-ray study of some sounds of Australian English. *Phonetica, 21*(3), 138-150.

[5] Wells, J. C. (1982). *Accents of English* (Vol. 1). Cambridge University Press.

[6] Cox, F. (2006). The acoustic characteristics of /hVd/ vowels in the speech of some Australian teenagers. *Australian Journal of Linguistics*, *26*(2), 147-179.

[7] Harrington, J., Cox, F., & Evans, Z. (1997). An acoustic phonetic study of broad, general, and cultivated Australian English vowels. *Australian Journal of Linguistics*, *17*(2), 155-184.

[8] Wood, S. (1979). A radiographic analysis of constriction location for vowels. *Journal of Phonetics, 7,* 25-43.

[9] Stevens, K. N. (1989). On the quantal nature of speech. *Journal of Phonetics*, *17*, 3-45.

[10] Dawson, K. M., Tiede, M. K., & Whalen, D. H. (2016). Methods for quantifying tongue shape and complexity using ultrasound imaging. *Clinical Linguistics & Phonetics*, *30*(3-5), 328-344.

Nasal aerodynamics and coarticulation in Bininj Kunwok: Smoothing Spline Analysis of Variance

Hywel Stoakes[1], *Janet Fletcher*[1], *Andrew Butcher*[2]

[1]The University of Melbourne, Australia
[2]Flinders University, Australia

hstoakes@unimelb.edu.au j.fletcher@unimelb.edu.au andy.butcher@flinders.edu.au

Abstract

Nasal phonemes are well represented within the lexicon of Bininj Kunwok.[1] This study examines intervocalic, word medial nasals and reports patterns of coarticulation using a Smoothing Spline Analysis of Variance (SSANOVA). This allows for detailed comparisons of peak nasal airflow across six female speakers of the language. Results show that in a VNV sequence there is very little anticipatory vowel nasalisation and greater carryover into a following vowel. The maximum peak nasal flow is delayed for coronals when compared to the onset of oral closure in the nasal, indicating a delayed velum opening gesture. The velar place of articulation is the exception to this pattern with some limited anticipatory nasalisation. The SSANOVA has shown to be an appropriate technique for quantifying these patterns and dynamic speech data in general.

Index Terms: nasals, SSANOVA, aerodynamics, Australian languages

1. Background

1.1. Nasals in Australian languages

Australian languages have sonorant-rich phoneme inventories with nasals at many places of articulation, matched with oral plosives. Many languages contain five [1], six [2] and sometimes seven [3] contrastive nasals, with additional laterals matched with coronals. Across the areal phylum, vowels are not documented as contrasting phonemically based on nasalisation [1]. Previous research across languages generally, shows that coarticulation makes segments less phonemically distinct and in order to keep phonemes phonologically contrastive coarticulation needs to be limited [4], [5]. In Australian languages, due to the large number of place of articulation contrasts, there is a phonological imperative to keep nasals perceptually separated from each other, yet the mechanisms behind this are still understudied [1].

In Australian languages anticipatory nasalisation is thought to be very tightly controlled. The phoneme inventories of Australian languages commonly contain many places of articulation which need to be distinguished acoustically. Phonetic nasalisation in vowels makes it more difficult to discriminate the place of articulation of following nasals [6]. This is because place of articulation cues are often marginal in nasals with acoustic cues found within the transitions between nasal and vowels most salient (see [7] for an overview). The perceptual consequences of these spectral concentrations is that formant transitions at the margins of the nasal rather than the low frequency nasal murmur convey the majority of the place of articulation information (e.g. [6] for American English, and [7] for Catalan). In order to preserve the place of articulation cues in nasals and surrounding segments, a delay in velum lowering would limit the confounding effect of vowel nasalisation on cues that are due to movements in the oral articulators [8]. The delay provides the maximum opportunity for the perception of theses transitional cues found within the speech spectrum within sonorants and vowels ([1], [9]).

In a related vein, previous studies of Australian languages show that Warlpiri [10] and Iwaidja [11], both allow temporal coproduction of an apical nasal with a dorsal stop showing only limited spatial modifications, particularly in apical nasals [9]. In Burarra, Gupapuyŋu, and Warlpiri, anticipatory vowel-consonant coarticulatary resistance exceeds that of carry-over coarticulation [12]. These results suggest that a coarticulatory gesture can be anticipated and controlled by the speaker and may be consciously resisted in order to keep phonemic categories distinct. The current study asks whether, in order to control the extent of anticipatory coarticulation, advanced planning motivated by a need to preserve place of articulation cues is needed to mitigate the masking or loss of crucial spectral or articulatory cues.

1.2. Smoothing Spline ANOVA

There has been increased interest in the quantitative analyses of dynamic speech data in recent years. This has led to renewed focus on the acoustic analysis of dynamic formants and fundamental frequency measured over time. The Smoothing Spline Analysis of Variance (SSANOVA) method, introduced by Gu [13] is one method of averaging complex time-series data. There are now a number of phonetic studies utilising this technique for a variety of dynamic speech data. The most prevalent to date have been the analysis of static ultrasound tongue splines [14],[15], [16], [17] and acoustic formant data ([18],[19], [20], [21]) which allow dynamic formant trajectories to be compared across speakers and words. Levels of speaker variation prove challenging to analyse and this technique is promising for comparing articulatory results across speakers.

In this study peak airflow is averaged using a similar spline smoothing algorithm, utilising functions contained within the *gss* package [22]. In order to apply an SSANOVA successfully to speech data, each segment must first be temporally normalised before subsequent statistical analyses are calculated (see Section 2.4). Each phoneme is considered individually with separate confidence intervals. The resulting plot, averaged across speakers, indicates the peak nasal airflow rate plotted across time separately for each of the nasal phonemes.

[1] There was a recent decision to standardise the orthography to use the Kunwinjku conventions, thus *Gun-wok* is now *Kunwok* (see http://bininjgunwok.org.au/information/orthography/)

2. Methods

2.1. Speakers and Materials

The recordings in were made with six female speakers of Bininj Kunwok (Kunwinjku variety) who repeated a list of disyllabic lexical items each containing intervocalic medial nasals. The lists were compiled by the first and third authors with reference to the Kuninjku Dictionary [23] and the Kunwinjku learners dictionary [24]. The list was then checked and revised by Murray Garde and by the first author in consultation with Bininj Kunwok speakers to ensure both semantic and phonological accuracy.

All words begin with a voiceless velar stop (except *bininj*) and each was uttered within the same carrier phrase (*yun yime X yimen Y*). Each speaker made three repetitions of the word list, although not all recordings were usable due to data capture errors, giving a total of 107 tokens (see Table 1).

Table 1: *Word list and number of tokens (n).*

Word:	*bininj*	*kamak*	*kangokme*	*kanjok*
Phonetic:	[ˈpɪnɪɲ]	[ˈkɐmɐk]	[kɐˈŋɔkmɛ]	[ˈ kɐɲɔk]
Gloss:	‘male’	‘good’	‘carry away’	kin-term
n	22	11	9	15
Word:	*karnubirr*	*kinga*	*kumoken*	*kunak*
Phonetic:	[ˈkɐɲˌʊbɪr]	[ˈkɪŋɐ]	[ˈ kʊmɔˌkɛn]	[ˈkʊnɐk]
Gloss:	f.w. mussel	‘crocodile’	‘f.w. crocodile’	‘fire’
n	8	13	13	16
Total:	*107*			

2.2. Aerodynamic Recordings

This study reports the results from a single peak nasal airflow channel (U_n measured in $\mathrm{cm}^3\,\mathrm{s}^{-1}$). Simultaneous peak oral airflow (U_o) was also recorded, although oral airflow data are not reported here. The multichannel articulatory recordings were gathered via Scicon R&D oral and nasal airflow masks with an in-built microphone connected to a Scicon R&D 916 capture device. The airflow acquisition hardware was controlled using the PCQuirer software (Version 7, Scicon R&D California, USA). Calibration was done before and after the equipment was moved to the field site.

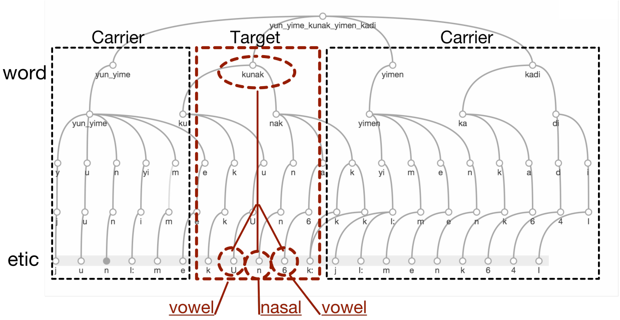


Figure 1: *A example of the hierarchy showing the target word within the carrier phrase*

2.3. Labelling and Querying

Labelling and segmentation was done within the Emu WebApp [25] and further analyses used the *EmuR* package [26] within the R programming environment [27]. The acoustic signal was used as the basis of segmentation and the determination of the nasal vowel boundary. The hierarchical querying architecture of Emu is essential in order to restrict the measurements of the target nasal to the word medial intervocalic position (V_1NV_2). The following code queries a hierarchy that has both word and phonetic (etic) tiers temporally linked within Emu (see Figure 1).[2]

```
#queries for a VNV sequence
require(emuR)
# data base first loaded using:
BGW_AE_N_2006 <- load_emuDB(databasepath)

VNV.seq <- emuR::query(BGW_AE_N_2006,
    "[[etic = vowel ->
      [etic = nstop & Medial(word,etic) = 1
             ˆ word =˜ .*]] ->
       etic = vowel]",
timeRefSegmentLevel = "etic",
resultType = "emuRsegs"
                            )
```

This gives an *R* vector (an *Emu* segment list) containing the results for a word medial sequence of a medial nasal surrounded by two vowels. This can be refined to remove the carrier phrase tokens. We then return an individual segment list for each of the items in turn by specifying the target segment (see the documentation for the *Emu Query Language Version 2* for details). This gives four parallel vectors, one containing data for *V1*, one containing data for *N*, one containing data for *V2* and one that encompasses the entire word, used in subsequent analysis. The airflow channels are then extracted using the `emuR::get_trackdata` function providing an *R* data frame.

2.4. Normalisation

We use a similar aerodynamic measurement methodology to that reported by [28] for French nasal sequences in that the nasal flow is averaged for all speakers for each individual phoneme and then the sequence is then reconstructed in temporal order. The airflow (U_n) averaging is achieved by first, time normalising the signal and subsequently averaging the airflow for each segment separately which gives an average peak flow over time (U_n) [28, pp 594–5]. This method shows the absolute timing of dynamic changes in airflow. The flow magnitude information, however–as it is an average across speakers–is less valuable. A smoothing spline ANOVA is then calculated (see Section 2.5 below for the method). Each token has had the zero-offset, adjusted as over the course of a recording session the zero flow level drifted either upward or downward. The minimum value in V_1 was measured and used as the zero value for the entire sequence which was then then used to normalise the airflow values in N and V_2.

2.5. SSANOVA

The process for calculating the SSANOVA closely follows the method introduced by Fruewald [19] who compared dynamic formant trajectories (F1 and F2). In the current study an SSANOVA (`gss::ssanova()`) is calculated using the *gss* package and subsequently the `stats::predict()` function which makes a prediction for each point based on the model (`fit`). The corresponding standard error is also calculated (`se.fit`). These

[2] Anonymised data can be accessed at http://hywel.github.io/data/df_VNV_V1.csv (1Mb)

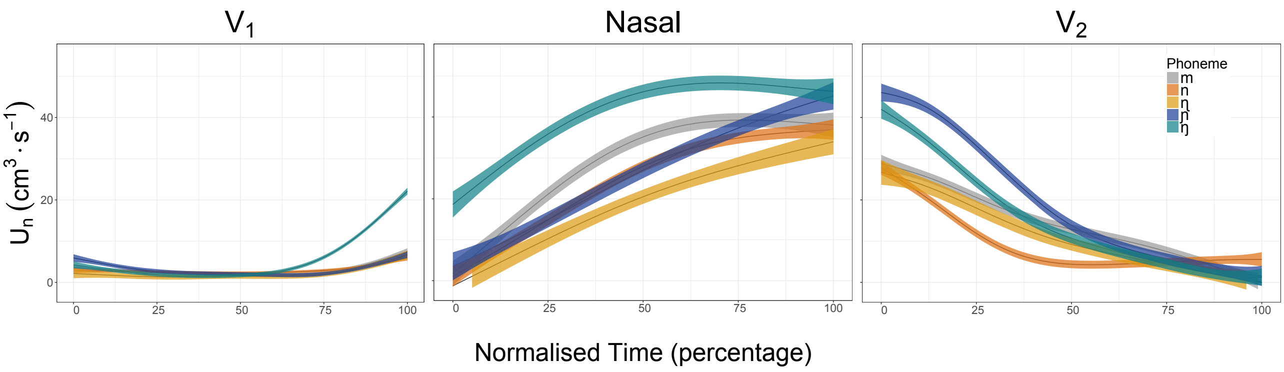


Figure 2: *An SSANOVA of average nasal airflow by normalised time in Vowel*$_1$*, Nasal, Vowel*$_2$ *sequences separated by phoneme*

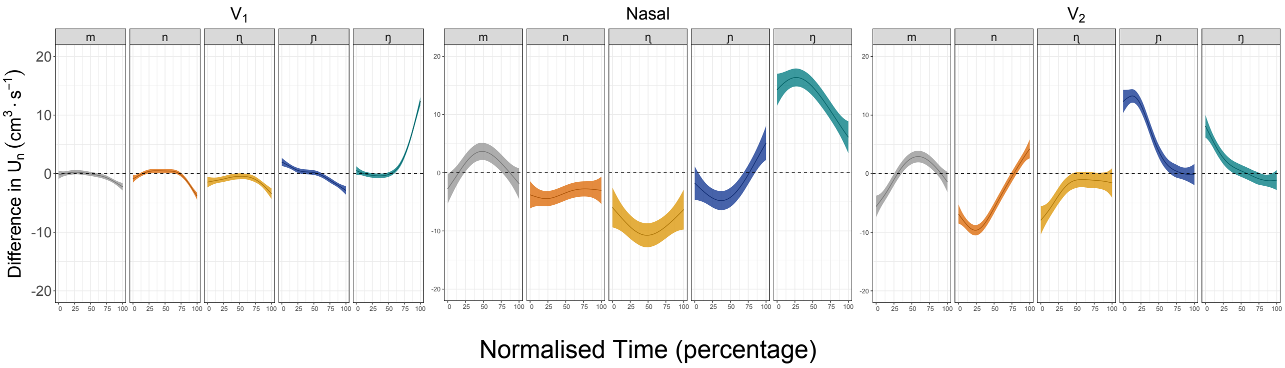


Figure 3: *The group plus interaction over normalised time between the phonemes (nasal:time) in Vowel*$_1$*, Nasal, Vowel*$_2$ *sequences*

are returned individually for each nasal phoneme (`nasal`) and results are then plotted using the *ggplot2* package. The *R* [27] code below produces a vector allowing plotting of the first panel (Vowel 1) shown in Figure 2.

```
df.VNV.V1 <-read.csv
          (file = "df.VNV.V1.csv")
require(gss)
Un.VNV.V1.model <-
  ssanova(data˜nasal + time + nasal:time,
          data = df.VNV.V1)
grid.VNV.V1 <-
  expand.grid(time = seq(0,1,length = 100),
  nasal = c("m","n","ɳ ","ɲ","ŋ"))
grid.VNV.V1$Un.Fit <-
  predict(Un.VNV.V1.model,
           data_n = grid.VNV.V1,
           se = T)$fit
grid.VNV.V1$Un.SE <-
  predict(Un.VNV.V1.model,
          data_n = grid.VNV.V1,
          se = T)$se.fit
```

This generates the data frame for the first panel of the plot in Figure 2. The group and time interaction is then shown in Figure 3 indicating the difference in airflow over time for each phoneme.

3. Results

3.1. Smoothed nasal airflow over time

The figures 2 and 3 report the results from the SSANOVA. Figure 2 shows the SSANOVA for each phoneme plotted across 100 sample points as a percentage. The standard error (se) is shown as a ribbon in Figure 2) and is calculated using a 95% Bayesian confidence interval. The discontinuities at the edges of the panels are due to the averaging of the U_n signal over time and minor perturbations in the flow signal. Figure 3 shows the group interaction between the phonemes over time (nasal:time). When the plots intersect the zero line it indicates that the phonemes are not significantly different at that timepoint.

Results show very little anticipatory nasalisation in a vowel preceding a word medial nasal for all phonemes except the velar. The velar nasal shows an increased peak nasal airflow starting at 65% of the initial vowel (V_1). During the peripheral nasals /m/ and /ŋ/ have peak nasal airflow that occurs before that of the other phonemes with /ŋ/ just after 25% into the nasal and /m/ just prior to 50%. The coronal consonants all have their maximum peak of nasalisation at the acoustic offset of the nasal (centre panel of figure 2). Figure 3 shows that for velars (/ŋ/)the difference in flow is greater in V_1 and N than the other phonemes. The palatal (/ɲ/) has a higher peak flow than the each of the other phonemes in the second vowel (V_2)indicating that it has the highest carryover nasalisation. This carryover effect may be due to the greater contact area of the laminal articulator meaning that coordination between oral closure and nasalisation is more difficult to maintain. In velars, velum lowering is less delayed because, unlike with coronal articulations, the velum needs to be lowered in order to make closure with the tongue dorsum during the articulation of the nasal itself.

4. Discussion

This study shows that anticipatory coarticulation of vowel nasalisation is tightly controlled in medial VN sequences by Bininj Kunwok speakers. These patterns of nasal flow are interpreted as evidence of delayed velum lowering during the pre-nasal vowel. It is clear from the results that carryover nasalisation is not controlled in the same manner and that the peak of nasalisation is at the offset of oral closure for the coronal nasals. The variation in the location of the peak nasal flow suggests that there are physical differences between the articulation of these phonemes although this is not thought to be at the level of awareness. The tight control of velum lowering may be used as a strategy to ensure that place of articulation information is phonetically retrievable in an environment that can obscure place of articulation cues. This equates very well with qualitative examinations of acoustic signals in Bininj Kunwok, suggesting that the SSANOVA technique is appropriate for the analyses of complex time-course data. Further work will look at the duration and timing of the both the opening and closing phase in the language.

5. Acknowledgements

The authors would like to thank the speakers from the Mamardawerre community, Deborah Loakes and Murray Garde, as well as two anonymous reviewers. All errors remain our own. This research is supported by the ARC Centre of Excellence for the Dynamics of Language.

6. References

[1] Butcher, A. R., “Consonant-salient phonologies and the ‘place-of-articulation imperative’,” in Speech Production: Models, Phonetic Processes and Techniques, Harrington, J. and Tabain, M., Eds., New York: Psychology Press, 2006, 187–210.

[2] Tabain, M. and Butcher, A. R., “Stop consonants in Yanyuwa and Yindjibarndi: Locus equation data,” Journal of Phonetics, 27 (1), 333–357, 1999.

[3] Breen, G., “The wonders of Arandic phonology,” in Forty Years On: Ken Hale and Australian Languages, Simpson, J., Nash, D., and Laughren, M., Eds., Canberra: Pacific Linguistics, 2001, 45–69.

[4] Manuel, S. Y., “The role of contrast in limiting vowel-to-vowel coarticulation in different languages,” The Journal of the Acoustical Society of America, 88 (3), 1286–1298, 1990.

[5] Scarborough, R., Zellou, G., Mirzayan, A., and Rood, D. S., “Phonetic and phonological patterns of nasality in Lakota vowels,” Journal of the International Phonetic Association, 45 (03), 289–309, 2015.

[6] Miller, G. and Nicely, P., “An analysis of perceptual confusions among some English consonants,” The Journal of the Acoustical Society of America, 27, 338, 1955.

[7] Recasens, D., “Place cues for nasal consonants with special reference to Catalan,” Journal of the Acoustical Society of America, 73 (4), 1346–1353, 1983.

[8] Butcher, A. R., “What speakers of Australian Aboriginal languages do with their velums and why: The phonetics of the nasal/oral contrast,” in XIVth International Conference of Phonetic Science, San Francisco, 1999.

[9] Fletcher, J., Butcher, A., Loakes, D., and Stoakes, H., “Aspects of nasal realization and the place of articulation imperative in Bininj Gun-wok,” in Proceedings of the 13th Speech Science and Technology Conference, Melbourne, ASSTA, 2010.

[10] Fletcher, J., Loakes, D., and Butcher, A., “Coarticulation in nasal and lateral clusters in Warlpiri,” in Annual Conference- International Speech Communication Association, ser. International Speech Communication Association, 1, 2009, 86–89.

[11] Fletcher, J. M., Butcher, A. R., Loakes, D., and Stoakes, H., “Coarticulation and consonant cluster production in Iwaidja,” The Journal of the Acoustical Society of America, 129 (4), 2452–2452, 2011.

[12] Graetzer, S., Fletcher, J., and Hajek, J., “Locus equations and coarticulation in three Australian languages,” The Journal of the Acoustical Society of America, 137 (2), 806–821, 2015.

[13] Gu, C. and Wahba, G., “Smoothing spline ANOVA with component-wise Bayesian “confidence intervals”,” Journal of Computational and Graphical Statistics, 2 (1), 97–117, 1993.

[14] Davidson, L., “Comparing tongue shapes from ultrasound imaging using smoothing spline analysis of variance,” The Journal of the Acoustical Society of America, 120 (1), 407–415, 2006.

[15] Billington, R., “‘Advanced tongue root’ in Lopit: Acoustic and ultrasound evidence,” in Proceedings of the 15th Australasian International Speech Science and Technology Conference, Hay, J. and Parnell, E., Eds., ASSTA, 2014.

[16] Heyne, M. and Derrick, D., “Using a radial ultrasound probes virtual origin to compute midsagittal smoothing splines in polar coordinates,” The Journal of the Acoustical Society of America, 138 (6), EL509–EL514, 2015. [Online]. Available: http://scitation.aip.org/content/asa/journal/jasa/138/6/10.1121/1.4937168.

[17] Mielke, J., “An ultrasound study of Canadian French rhotic vowels with polar smoothing spline comparisons,” The Journal of the Acoustical Society of America, 137 (5), 2858–2869, 2015.

[18] De Decker, P. and Nycz, J., “A new way of analyzing vowels: Comparing formant contours using smoothing spline ANOVA,” in Proceedings of thet the 35th NWAV Conference, 2006.

[19] Fruehwald, J., “SS ANOVA,” Academia.edu, 2010. [Online]. Available: https://www.academia.edu/268789/SS_ANOVA.

[20] Haddican, B., Foulkes, P., Hughes, V., and Richards, H., “Interaction of social and linguistic constraints on two vowel changes in northern England,” Language Variation and Change, 25 (03), 371–403, 2013.

[21] Docherty, G., Gonzalez, S., and Mitchell, N., “Static vs. dynamic perspectives on the realization of vowel nuclei in West Australian English,” in 18th International Congress of Phonetic Sciences. Glasgow, Scotland, 2015.

[22] Gu, C., “Smoothing spline ANOVA models: R package gss,” Journal of Statistical Software, 58 (5), 1–25, 2014. [Online]. Available: http://www.jstatsoft.org/v58/i05/.

[23] Garde, M., Bininj Kunwok Dictionary. ANU, forthcoming.

[24] Manakgu, A. and Etherington, S., Basic Kunwinjku Dictionary: A simplified English-Kunwinjku and Kunwinjku-English dictionary. The Kunwinjku Language Centre, 1996.

[25] Winkelmann, R. and Raess, G., “Introducing a web application for labeling, visualizing speech and correcting derived speech signals,” in Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC’14), Chair), N. C. (, Choukri, K., Declerck, T., Loftsson, H., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., and Piperidis, S., Eds., Reykjavik, Iceland: European Language Resources Association (ELRA), May 2014, 26–31, ISBN: 978-2-9517408-8-4.

[26] Winkelmann, R., Jaensch, K., Cassidy, S., and Harrington, J., EmuR: Main package of the Emu Speech Database Management System, R package version 0.1.8.9001, IPS Munich. [Online]. Available: https://github.com/IPS-LMU/emuR.

[27] R Core Team, R: A language and environment for statistical computing, R Foundation for Statistical Computing, Vienna, Austria, 2016. [Online]. Available: https://www.R-project.org/.

[28] Delvaux, V., Demolin, D., Harmegnies, B., and Soquet, A., “The aerodynamics of nasalization in French,” Journal of Phonetics, 36 (4), 578–606, 2008.

Preliminary Investigations into the Australian English Articulatory Vowel Space

Louise Ratko[1]*, Michael Proctor*[1]*, Felicity Cox*[1]*, Sean Veld*[2]

[1]Department of Linguistics, Macquarie University, Australia
[2]Department of Cognitive Science, Macquarie University, Australia

{louise.ratko, michael.proctor, felicity.cox}@mq.edu.au, sean.veld@students.mq.edu.au

Abstract

Articulation of vowels produced by a single speaker of Australian English in CVC contexts was examined using Electromagnetic Articulography. Dorsal articulatory activity for each vowel was compared by tracking the midsagittal trajectories of the tongue body. Articulatory targets were determined and a companion articulatory vowel space constructed. Comparison of dorsal trajectories of vowel pairs /ɐː-ɐ/, /eː-e/, /oː-ɔ/ confirms a close articulatory relationship between long-short pairs that has previously only been examined in the acoustic domain. Long vowels were characterised by greater excursion from the centre of the midsagittal articulatory space, compared to their short equivalents.

Index Terms: Vowel Production, Australian English, Articulatory Phonetics, Electromagnetic Articulography, Kinematics

1. Introduction

The relationship between articulatory and acoustic targets for vowels is complex and remains imperfectly understood [1]. It is still not clear to what extent vowel contrasts are defined in terms of acoustic [2,3] or articulatory [4,5] goals of production, or both [5,6]. A limiting factor in our understanding of these issues is the relative paucity of articulatory data available, particularly on Australian English (AusE) vowel production.

The explanation for this scarcity is twofold; the acoustic properties of vowels position them as ideal candidates for spectrographic analysis [1,2,4]. This coupled with the historical difficulty of obtaining readily analysable articulatory data has meant that progress in vowel articulation research has lagged behind work in the acoustic domain even as technology improves.

While ultrasound and x-ray have been used for vowel analysis with a focus on tongue shape [7,8], kinematic data from Electromagnetic Articulography (EMA) provides high resolution tracking of tongue midsagittal displacement across vertical, horizontal and lateral dimensions in relation to both the participant's occlusal plane and hard palate. EMA is particularly useful for tracking the velocity and acceleration of lingual regions [9]. These data are important for informing models of articulatory kinematics in relation to vowel production [10].

In this study we examine the production of AusE monophthongs in a CVC context, by a single speaker. As a preliminary investigation, we explore the viability of characterising AusE vocalic targets using articulatory kinematics and present some initial observations about patterns of dorsal articulation to complement existing acoustic findings on AusE vowels.

1.1. Articulatory vowel space

Traditionally in the acoustic analysis of monophthongs, a single point is chosen, usually midway through the vowel, to represent the vowel's acoustic target [1, 11]. The values of (at least) the first two formants may be extracted at the target to construct a vowel space diagram. The F1 (high-low) and F2 (front-back) dimensions, when plotted on the Y and X axes respectively, indicate each vowel's relative position in terms of phonetic height and fronting and are often sufficient to differentiate vowels in the inventories of many of the world's languages [1].

In AusE the F1 x F2 vowel space is sufficient to differentiate all but four of the 12 AusE stressed monophthongs [11, 12], with /ɐː, ɐ/ '*Bart'* vs *'but'* (and possibly /eː, e/ *'bared'* vs *'bed'*) differing only in duration [9]. The acoustic vowel space is a useful tool for comparing vowels in a range of applications such as the investigation of differences between languages, vowel change, differences between social groups, and phonetic or prosodic contexts [11, 12, 13, 14].

Watson et al. [13] have demonstrated the limitations of the formant relationships described by the acoustic vowel space in their kinematic and acoustic study of the New Zealand English (NZE) high front vowel in 'hid'. Acoustic analysis has indicated that the target of this vowel has both retracted and lowered over time. Yet this was not reflected in the articulatory position of the tongue dorsum during target production, with a high tongue position maintained by all five of their participants.

To reconcile the discrepancy between acoustic and articulatory evidence, and unravel the complex relationship between vowel acoustics and vowel articulation, a methodological approach to the creation of an articulatory vowel space must be established. The present study aims to determine whether the articulatory parameters of tongue dorsum height and tongue dorsum fronting are useful in differentiating the monophthongs of AusE, and whether the relative dorsal positions of these vowels reflect their relative positions in the acoustic vowel space.

1.2. Dynamic articulatory trajectories

Although it is possible for AusE vowels to be differentiated by their acoustic target and duration [11], many intrinsic properties of vowels are dynamic in nature. The presence and length of acoustic onglides and offglides [11, 12] and formant transitions from the preceding and into the following consonant may improve vowel differentiation rates, even for monophthongal vowels [11].

This is not the case for all AusE monophthongs: the long-short vowel pairs /ɐː-ɐ/ and /eː-e/ have been shown to differ only in duration, with formant trajectories and acoustic target values near identical [12]. This is in contrast to other long-short vowel combinations in AusE such as /oː- ɔ/, which differ in spectral, as well as temporal properties.

Thus to further explore dynamic characteristics of these vowel pairs we will examine midsagittal tongue dorsum trajectories during articulation of /ɐː-ɐ/ (in *Bart* and *but*) and /eː-e/ (in *bared* and *bet*) and compare these to the dorsal trajectories of the spectrally distinct vowel pair /oː-ɔ/ (in *bought* and *pot*). It is predicted that within the long/short vowel pairs /ɐː-ɐ/ and /eː-e/ similar articulatory trajectory shapes and articulatory targets will be found consistent with prior acoustic descriptions of these vowel pairs [12]. We expect the items within the vowel pair /oː-ɔ/ to show distinct dorsal trajectories and distinct target locations as suggested by their acoustic descriptions [12].

2. Method

2.1. Participant and speech material

A native speaker (Sydney, male, 45 y.o.) produced the 18 stressed vowels of AusE across a variety of phonetic contexts including hVd, bVt and tVn. bVt tokens were chosen for preliminary analysis as they provided two clear articulatory landmarks; the labial /b/ gesture and the coronal /t/ gesture, between which articulatory trajectories could be compared across tokens. /eː/ was only elicited in bVd context. All vowels were elicited from common English words.

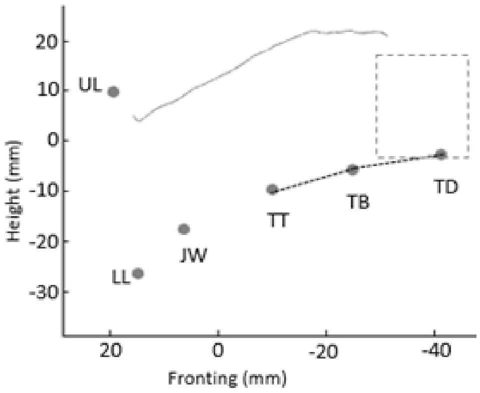


Figure 1: ***Configuration of articulators during /ɐː/ production****. Sensors attached to: Upper Lip (UL); Lower Lip (LL), Jaw (JW), Tongue Tip (TT), Tongue blade (TB), and Tongue dorsum (TD). Solid line indicates midsagittal palate trace. Vertical origin located at occlusal plane. Dashed line between TT-TB-TD sensors approximates midsagittal tongue line. Dashed box indicates limits of excursion of TD sensor values during vowel articulation.*

2.2. Data Acquisition

Articulatory data were acquired using an NDI Wave system sampling each sensor at a rate of 100 Hz. Nine sensors were attached to the participant. Three (nasion, left and right mastoid) were used to correct for head movement, two sensors (upper, lower lip) tracked labial aperture, and a sensor attached below the lower incisors tracked jaw movement. Three sensors were affixed to the midsagittal line of the tongue at the tongue tip (TT) 25 mm from participant's anatomical tongue tip, tongue blade (TB) 20 mm posterior to TT sensor and tongue dorsum (TD) 35 mm posterior to TT sensor (Fig.1). The occlusal plane was located with a bite trial, and the midline of the palate was traced with a custom 6D palate probe (Northern Digital Inc.).

2.3. Data analysis

2.3.1. Acoustic/ and articulatory vowel space

Acoustic vowel targets were extracted from spectrogram using Praat [15] and plotted in a traditional F1 x F2 vowel space (Fig. 3) based on criteria outlined in [12].

Articulatory data were analysed in MView [16]. For each token, an automatic gesture labelling procedure was used to locate the dorsal articulatory target of the nuclear vowel, from the tangential velocity profile of the TD sensor in each vocalic interval [17]. The articulatory target was the point at which the tongue dorsum sensor reached a point of minimum velocity. Tongue dorsum height was calculated in relation to the participant's occlusal plane; tongue dorsum fronting was determined in relation to the point of maxillary occlusion in line with [6, 7, 14].

2.3.2. Dorsal Trajectory

A comparable interval for vocalic analysis was identified in each token by locating articulatory landmarks in the marginal consonants. The vocalic analysis interval was bounded by the labial onset release and the closure of the coronal coda (Fig. 2).

The midsagittal location of the dorsal sensor was tracked through the interval defined between the onset and coda consonants in each token (Fig.2). For each vowel, this produced a 2-dimensional signal indicating the location of the top of the tongue dorsum in the midsagittal plane at 10 ms intervals which was used to plot dynamic midsagittal dorsal articulatory trajectories.

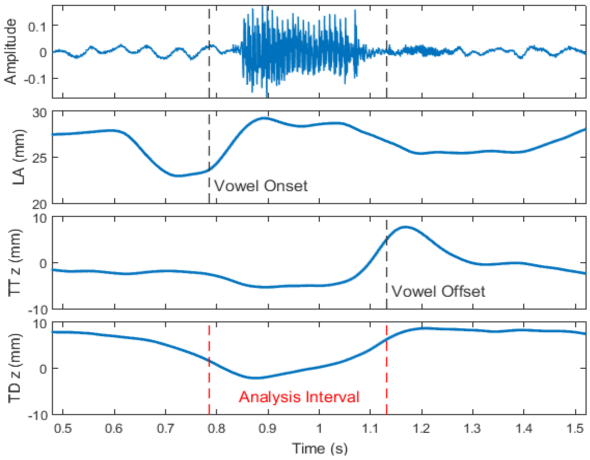


Figure 2: ***Method of location of interval of analysis****. Top: speech waveform for utterance 'bert' /bɜːt/; 2nd row: time aligned labial aperture (UL-LL mm); 3rd row: vertical location of TT sensor (mm); 4th row: vertical location of TD sensor (mm). Start of vocalic analysis interval: release of /b/ closure gesture; End of interval: onset of /t/ closure.*

3. Results

3.1. Acoustic vowel space

F2 and F1 for the participant's vowels are plotted in Fig.3. Comparison with vowel data reported in the Sydney AusTalk corpus [18] shows that the majority of the participant's tokens were within two standard deviations of reported means, with slightly lower F1 values reported for /æ, ɐ, ɐː/, and slightly lower F2 values for /ɔ/.

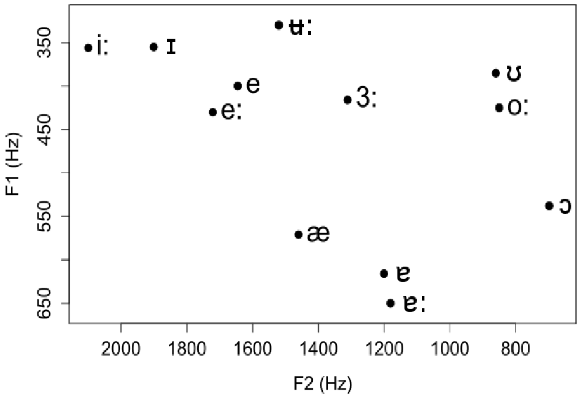


Figure 3: ***Acoustic Vowel Space.*** *First and second formants measured at point of maximum acoustic stability in each vocalic interval.*

3.2. Articulatory Vowel Space

Dorsal targets in the midsagittal plane for each vowel in the corpus are compared in Fig. 4. The relative configuration of articulatory targets is in general agreement with the acoustic spacing of the vowels produced by this speaker, with some notable exceptions. /ʉː/ is much more fronted in the articulatory space compared to its acoustic target, /oː/ is acoustically closer to /ʊ/ yet articulatorily closer to /ɔ/, and /ɜː/ has a low front dorsal target, yet a high mid acoustic target.

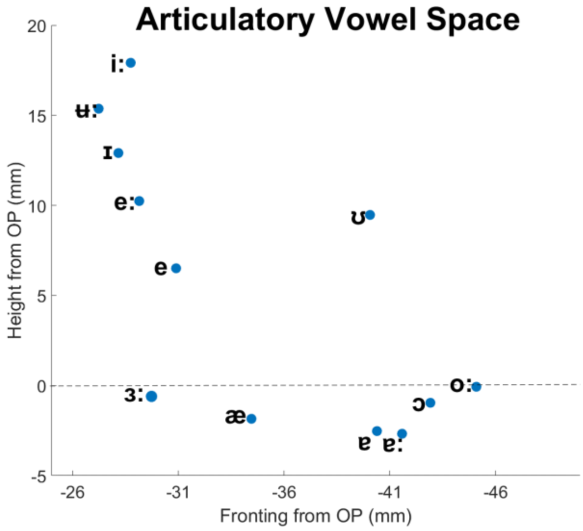


Figure 4: ***x-y coordinates of TD sensor at articulatory target of nuclear gesture in bVt/d tokens.*** *Dorsal height (mm) indicated with respect to Occlusal Plane (OP)*

3.3. Dynamic articulatory trajectories

Midsagittal TD trajectories for long-short vowel pairs /ɐː-ɐ/, /eː-e/ and /oː-ɔ/ produced in bVt/d contexts are illustrated in Figs. 5-7. The short vowels in the /ɐː-ɐ/ and /eː-e/ pairs were produced with similar trajectories and direction paths compared to their long counterparts. Short vowels /ɐ/ and /e/ exhibit a more centralised trajectory than their long equivalents, which show more peripheral excursion. Similar dorsal trajectories, with the same orientation are observed for back vowels /oː-ɔ/ (Fig. 7), which share similar spatial targets in the midsagittal plane, despite their considerable differences in F1 and F2 frequencies (Fig. 3).

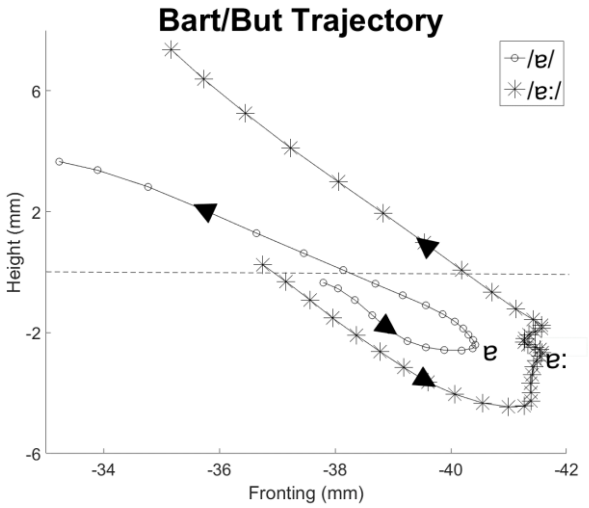


Figure 5: ***Midsagittal dorsal trajectories for long-short low vowels /ɐː-ɐ/.*** *IPA annotations indicate vowel target location. Sensor locations sampled at 10 ms intervals.*

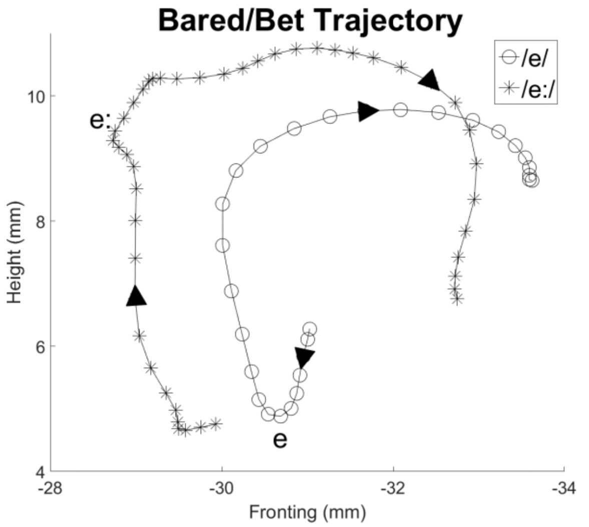


Figure 6: ***Midsagittal dorsal trajectories for long-short front vowels /eː-e/.*** *IPA annotations indicate vowel target locations. Sensor locations sampled at 10 ms intervals.*

4. Discussion

The vowel targets produced by this speaker were found to be broadly consistent with previous articulatory descriptions of AusE vowels [13]. Analysis of the relative distributions of

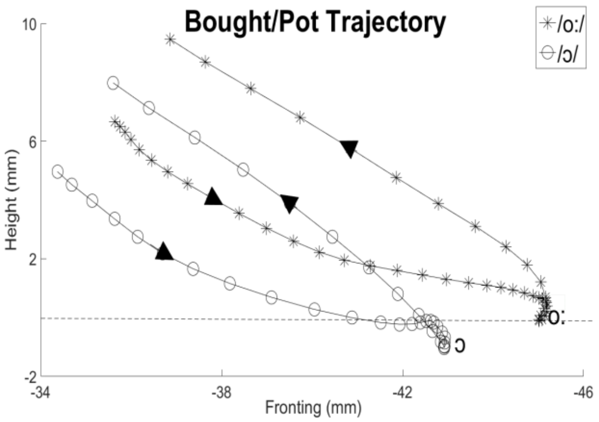


Figure 7: ***Midsagittal dorsal trajectories for long-short back vowels /oː-ɔ/.*** *IPA annotations indicate vowel target locations. Sensor locations sampled at 10 ms intervals.*

vowels in the midsagittal plane suggests that the articulatory dimensions of tongue dorsum height and dorsal fronting are sufficient to differentiate the majority of AusE monophthongs.

For /ɜː/ and /ʉː/, a discrepancy between the acoustic dimension of vowel fronting and the articulatory dimension of tongue dorsum fronting was observed, with a greater degree of dorsal fronting relative to the other vowels than the acoustic analysis of this speaker's second formant frequency suggests. A lower than expected dorsal height was also observed for back vowels /ʊ/ and /oː/, compared to the acoustic analysis of first formant targets for these vowels. This placed /oː/ in a similar dorsal articulatory space to /ɔ/ despite the acoustic similarity of /oː/ and /ʊ/.

These data suggest that a comprehensive characterisation of the AusE vowel space requires more articulatory data, as differences in dorsal posture are not always predictable from the relative acoustic properties of vowels. We acknowledge that analyses restricted to the tongue dorsum alone may fail to capture important parasagittal differences of articulation.

When we examine the dorsal target locations of the back-rounded vowels, it is clear that there is a discrepancy with the acoustic vowel space. Rounded vowels appear to be in relatively different positions in the acoustic space compared to the articulatory space. This is particularly the case for /oː/, /ɜː/ and /ʉː/. It is possible that differing degrees of lip-rounding could impact our perception of vowel height and mask quite different underlying dorsal articulations, in line with observations found in NZE [13].

Comparisons of /eː-e/, /ɐː-ɐ/ and /oː-ɔ/ reveal that the dorsal trajectories of the vowels within each pair show a close articulatory relationship. Articulatory targets of /e/ and /ɐ/ were found to be more centralised, consistent with the characterisation of these short vowels as undershot realisations of their more peripheral, longer equivalents [12]. This dorsal centralisation was not observed in production of /ɔ/, suggesting that /oː/ and /ɔ/ have distinct articulatory and acoustic targets. The similarity in dorsal trajectories observed for /oː-ɔ/ is unexpected, given that these vowels are typically realised with distinct acoustic characteristics in AusE [12].

As the realisation of /eː-e/, especially their length, may have been influenced by the difference in voicing of the following consonant (t/d), further investigation is required to examine the relationships between short and long vowels.

Extension of this research will involve the recruitment of more speakers, tokens and phonetic contexts. The investigation of articulatory parameters such as lip rounding and jaw height on AusE vowel acoustics is also warranted.

The creation of a methodological approach to kinematic vowel data analysis will provide a mechanism for fully understanding the complex relationship between vowel articulation and vowel acoustics, and offers an exciting opportunity for future researchers.

5. Conclusion

This study provides a preliminary characterization of the kinematics of dorsal articulation in Australian English vowel production. These data demonstrate that articulatory and acoustic vowel targets share a complex but close relationship, yet independent of acoustic representations, the articulatory dimensions of tongue dorsum height and tongue dorsum fronting are sufficient in differentiating AusE monophthongs in this speaker. Midsagittal dorsal trajectories revealed that long-short vowel pairs /ɐː-ɐ/, /eː-e/, /oː-ɔ/, previously identified as acoustically similar in AusE, also share a close articulatory relationship for this speaker. More data will be required to better understand the temporal and spatial mapping between acoustic and articulatory targets in Australian English vowel production.

6. References

[1] Ladefoged, P. & Johnson, K. "A Course in Phonetics". 7th ed., Boston: Cengage, 2015.

[2] Nearey, T. M. "Phonetic features for vowels". Bloomington: Indiana University Linguistics Club, Indiana, 1978.

[3] Perkell, J.S., Matthies, M.L., et al. "Speech motor control: Acoustic goals, saturation effects, auditory feedback and internal models". Speech Comm, 22:227–250, 1997.

[4] Abercrombie, D. "Elements of General Phonetics". Edinburgh: Edinburgh University Press; 1967.

[5] Lindblom, B. & Sundberg, J. "Acoustical consequences of lip, tongue, jaw and larynx movement" J. Acoust. Soc. Am, 50: 1166-1179, 1971.

[6] Noiray, A., Iskarous, K. & Whalen, D. "Variability in English vowels is comparable in articulation and acoustics" LabPhon, 5: 271-288, 2014.

[7] Hashi M., Westbury, J. R & Honda, K. "Vowel posture normalization" J. Acoust. Soc. Am., 104: 2426-2437, 1998

[8] Stone, M., Shawker, T. H., Talbot, T. L., & Rich, A. H. "Cross-sectional tongue shape during the production of vowels". J. Acoust. Soc. Am, 83(4): 1586-1596, 1988

[9] Perkell, J., Cohen, M., Svirsky, M., Matthies, M., Garabieta, I. & Jackson, M. "Electromagnetic midsagittal articulometer systems for transducing speech articulatory movements". J. Acoust. Soc. Am., 92: 3078-3096, 1992.

[10] Saltzman. E. & Munhall, K. "A dynamical approach to gestural patterning in speech production", Ecological Psychology, 1: 333-382, 1989.

[11] Watson, C. & Harrington, J. "Acoustic evidence for dynamic formant trajectories in Australian English vowels". J. Acoust. Soc. Am., 106: 459-468, 1999.

[12] Cox, F. "The Acoustic Characteristics of /hVd/ Vowels in the Speech of Some Australian Teenagers". AJL, 26: 147-179, 2006.

[13] Watson, C., Harrington, J & Palethorpe, S. "A kinematic analysis of New Zealand and Australian English vowel spaces". In ICSLP, 1998.

[14] Tabain, M., Perrier, P. & Savariaux, C. "A kinematic study of prosodic boundary effects on /i:/ articulation in French" in ICPhS, 2003.

[15] P. Boersma, and D. Weenink, *Praat: Doing phonetics by computer* (Version 5.4) [Computer program], 2009.

[16] Tiede, M. "MVIEW: software for visualization and analysis of concurrently recorded movement data, Haskins Laboratory, 2005

[17] Gafos, A., Kirov, C. & Shaw, J. "Guidelines for using Mview", 2010

[18] Cox, F. "Formant Frequencies and Durations for /hVd/ vowels. Department of Linguistics, Macquarie University, Sydney, NSW, 2015.

The early bilingual influence on speech and music processing

Liquan Liu

School of Social Sciences and Psychology, Western Sydney University, Australia

l.liu@westernsydney.edu.au

Abstract

Previous studies report incongruent findings whether bilingual infants face delays when perceiving native speech contrasts compared to monolinguals. We present three experiments targeting monolingual and bilingual infant vowel, linguistic pitch, and musical pitch perception in the first year after birth. Bilingual infants outperformed their monolingual peers in each experiment, contrasting previous findings. We propose a heightened acoustic sensitivity hypothesis: facing a complex language environment, bilingual infants pay more attention to input acoustic details than monolinguals crossing linguistic and musical domains.

Index Terms: infant, bilingualism, perception, language, music

1. Introduction

Human experience to sounds and languages begins before birth. Speech perception is essential for the mastery of an infant's native sound system and subsequently her native language. In the first year after birth, infants tune in from their initial sensitivities towards the patterns from their ambient environment, a process known as perceptual attunement [1]. This process is not restricted to the linguistic domain but applies to other fields such as musical [2] and face [3] perception. Some experiments have shown a delay in the vocabulary comprehension and production of a single language [4] among infants growing up learning two languages. Though under debate [5,6], discrepancies lead to the question whether the pace of sound acquisition may be delayed among bilingual infants. Since native sounds and words are acquired in a parallel fashion, the delay of one domain may affect the acquisition of the other.

Debates in bilingual speech perception begin with the study reporting the perception of a native vowel contrast during the language-specific perceptual attunement process. Spanish-Catalan bilingual infants of 8 months fail to discriminate Catalan-specific /e/-/ɛ/ and Catalan/Spanish /o/-/u/ contrasts [7,8]. Nevertheless, other studies have found that monolingual and bilingual infants are both sensitive to native contrasts [9,10]. Several accounts have been proposed for the observed bilingual delay, including, but not limited to, acoustic properties and salience, input frequency and distributional properties, rhythmic similarity or segmental variation (cognate words) between languages, phonetic space (the density of phonetic categories), segmental processing differences, task effects (tokens in use, number of talkers, paradigm, etc.), and social-indexical factors [7-10].

Despite the debates on the similarities and differences between monolingual and bilingual infants along the language developmental trajectory, recent studies illustrate advantages among bilingual infants, such as inhibition control [11] and attentional factors [12], in the cognitive domain. These advantages are attributed to the bilingual environment, even though no clear advantage has been shown in the linguistic domain for bilingual infants. The existence of such potential advantage is worth exploring.

Just like the language-specific perceptual attunement, infants present initial sensitivity to music [13], followed by the perceptual tuning process towards their native music conventions [2]. No previous research has tested the influence of bilingualism on music perception. We propose the following research question: do monolingual and bilingual infants differ in the discrimination of linguistic and non-linguistic distinctions between acoustic stimuli in the first year of life? To answer these questions, we tested infants on a native vowel, a non-native linguistic pitch, and a non-linguistic violin contrast before 12 months after birth.

2. Experiments

2.1. Exp.1 – Vowel perception

As the debate of speech perception between monolingual and bilingual infants begins with vowel perception, a native vowel contrast was selected in Exp.1.

2.1.1. *Participants*

Seventy-four 8-9-month-old Dutch monolingual and bilingual infants participated in the study. Data from 16 participants were excluded from analysis, the reasons being: fussiness (6), unable to habituate (1), inattentive (4), looking time (LT) less than 2 seconds for both trials in the test phase (2), and individual average LT for each sound category in the test phase more than 2 SD from the mean (3). Data from 58 participants were included for analysis, with 29 infants per language condition.

2.1.2. *Stimuli*

The Dutch /I/ and /i/ (e.g., rit 'ride' vs. riet 'reed') vowels were selected. The two vowels differ in the spectrum (first (F1) and second (F2) formant) but not duration [14,15]. This differs from the English /i:/-/i/ distinction in which both spectrum and duration differ. We hypothesized that limited acoustic cues in the Dutch contrast may increase infants' perceptual and learning difficulties. The syllables /bIp/ and /bip/ spoken by a female Dutch speaker were recorded in a sound-isolated booth with a DAT Tascam DA-40 recorder and a Sennheiser ME-64 microphone. The voice onset time values of syllable onsets and offsets /b/ and /p/ were set around 72ms and 1ms, respectively. The other natural properties of the contrast were maintained. The F1 and F2 values of the contrast are 409 and 2280 Hz for /I/ and 370 and 2597 Hz for /i/.

2.1.3. *Procedure*

Infants went through habituation and test phases. The habituation phase consisted of randomly presented tokens from one category. One trial ended when infants looked away for more than two seconds, and then the next trial began. The habituation criterion was set as infants' mean looking time of three consecutive trials dropping below 65% of the first three trials in the habituation phase. When habituated, the test phase started and infants heard two trials from the other category. Discrimination was indicated by a looking time recovery hearing the new stimuli. The stimuli presentation order between the two phases was counter-balanced.

During the experiment, infants sat on their parents' lap facing the screen, with the camera in front of them. Parents were blind to the testing stimuli and heard music from a headphone. No visual or audio interference could be observed in the test booth other than the stimuli used in the test. The test was conducted by a computer program [16]. A tester observed the experiments through a closed TV circuit and used a button box to record infants' looking times in the testing room adjacent to the testing booth. The inter-stimulus interval (ISI) was set as 1sec in all phases. If infants' looking time per trial was less than 2sec, the trial was considered ineffective and was excluded from the analysis [17].

2.1.4. *Results*

A Repeated Measures analysis of variance (RM ANOVA) was conducted with the mean of infant recovery looking time between the test trials and the end of habituation trials looking time as the within-subject variables and 2-level (monolingual vs. bilingual) language background as the between-subject factor. The effect of the phase change was significant, F(1,56) = 7.202, p = .010, η2 = .114. The interaction between language background and the phase change was also significant, F(1,56) = 6.784, p = .012, η2 = .108. Splitting the data by language background, paired samples t-test shows that the phase change was significant for bilingual, t(28) = -3.041, p = .005, but not monolingual group, t(28) = -0.080, p = .937. Hence, bilingual but not monolingual infants discriminated the contrast (Figure 1).

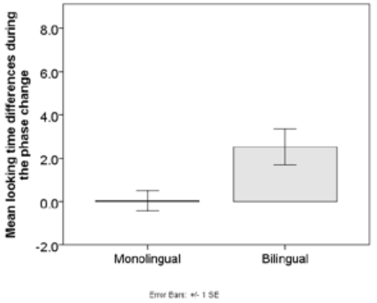


Figure 1: *Mean looking time differences in the phase change*

2.2. Exp.2 – Linguistic pitch perception

Exp.1 showed unexpected results that may be influenced by the sound categories of the native languages. To avoid such potential influence, we tested a linguistic pitch contrast novel to monolingual and bilingual infants in Exp.2.

2.2.1. *Participants*

Sixty-eight 11-12-month-old Dutch monolingual and (non-tone language learning) bilingual infants participated in the study. All bilingual infants were exposed to Dutch and another non-tone or pitch accent language. Data from 12 participants were excluded from the analysis for the following reasons: fussiness (1), unable to habituate (2), LT less than 2 seconds for both trials in the test phase (5), and individual average LT for each sound category in the test phase more than 2 SD from the mean (4). Data from 56 participants were included for analysis, with 28 infants per language condition.

2.2.2. *Stimuli*

The high-level (T1) and high-falling (T4) Mandarin Chinese tones were selected to create the stimuli. The tone-bearing syllable was /ta/. The productions of a Mandarin female speaker were recorded using the same device as in Exp.1. The natural Mandarin T1-T4 pair was further manipulated to avoid potential ceiling performance in non-native perception [18]. Considering the role of acoustic salience in non-native tone perception, the pitch distance between T1 and T4 was reduced to two fundamental frequency (F0) values occurring at 3/8 and 3/4 of the pitch distance of the original contrast, respectively, by introducing four interpolation points along the pitch contours (at 0%, 33%, 67% and 100%, see Fig.4). The new contrast shares the same acoustic properties with the T1-T4 contrast except for featuring a narrower distance between the pitch contours. The acoustic salience of this phonetic contrast is weakened by a pure manipulation of F0 (Contrast B, Figure 2).

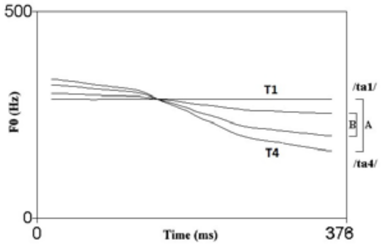


Figure 2: *The reduced T1-T4 [B] contrast shrunk from T1-T4 [A] to reduce the acoustic salience.*

2.2.3. *Procedure*

The same procedure as in Exp.1 was adopted.

2.2.4. *Results*

An RM ANOVA was conducted with the same within- and between-subject variables as Exp.1. The effect of the phase change was significant, *F*(1,54) = 4.976, *p* = .030, *η*2 = .084. The interaction between language background and the phase change was also significant, *F*(1,54) = 6.126, *p* = .016, *η*2 = .102. Splitting the data by language background, paired samples t-test shows that the phase change was significant for bilingual, *t*(27) = -2.655, *p* = .013, but not monolingual group, *t*(27) = 0.263, *p* = .794. Hence, bilingual but not monolingual infants discriminated the contrast (Figure 3).

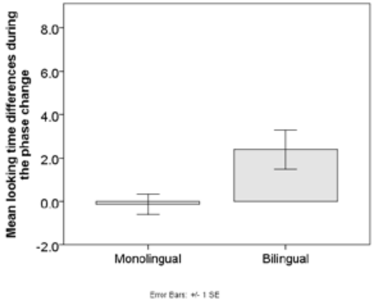


Figure 3: *Mean looking time differences in the phase change*

2.3. Exp.3 – Musical pitch perception

Exps.1 and 2 tested two contrasts in language. Results showed a bilingual perceptual advantage. In Exp.3, we explore monolingual and bilingual infant perception of a violin contrast in order to test the domain specificity of the effect observed in previous experiments.

2.3.1. *Participants*

Forty-eight 8-9-month-old Dutch monolingual and (non-tone language learning) bilingual infants participated in the study. Data from 12 participants were excluded from the analysis, the reasons being: fussiness (3), crying (3), inattentiveness (2), unable to habituate (1), and tone or pitch accent language exposure after birth (3). Eventually, data from 36 participants were included for analysis, with 18 infants per language condition.

2.3.2. *Stimuli*

To ensure the cross-domain comparison, the musical (violin) tonal stimuli were generated from the same contrast used in Exp.2. The F0 tiers of the contrast in Exp.2 were extracted and replaced the F0 tiers of a violin tone, creating novel violin stimuli. The violin contrast shared the exact same pitch contour as the tonal contrast in Exp.2 but differed in timber.

2.3.3. *Procedure*

The same procedure as in Exp.1 was adopted.

2.3.4. *Results*

An RM ANOVA was conducted with the same within- and between-subject variables as Exp.1. The effect of the phase change was significant, F(1,34) = 4.371, p = .044, η2 = .114. The interaction between language background and the phase change was also significant, F(1,34) = 4.565, p = .040, η2 = .118. Splitting the data by language background, paired samples t-test shows that the phase change was significant for bilingual, t(17) = -2.274, p = .036, but not monolingual group, t(17) = 0.062, p = .951. Hence, bilingual but not monolingual infants discriminated the contrast (Figure 4).

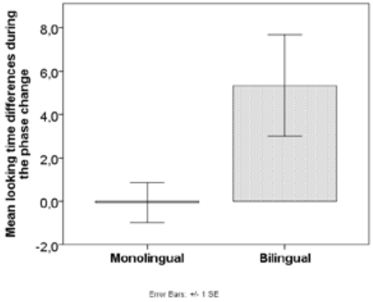


Figure 4: *Mean looking time differences in the phase change*

3. Discussion

The current study tested three contrasts between monolingual and bilingual infants in the first year after birth. Results showed differences between monolingual and bilingual infants and similarities across experiments. In Exp.1, monolingual infants did not discriminate the native vowel contrast. This finding differs from major language-specific perceptual attunement findings showing that sensitivity to native contrasts is maintained and improved in infancy [19], but conforms to some other studies illustrating the perceptual difficulty in some native contrasts [20]. Perception of native and non-native speech contrasts may be elastic [21], influenced by multiple factors including but not limited to frequency and age of exposure, as well as acoustic salience [22]. Contrasts with high acoustic salience can be discriminated across ages regardless of whether they are presented in the native sound inventories [23]; whereas contrasts with low acoustic salience (e.g., Exp.1) may initially be hard to discriminate, and infants' perception improves with their native linguistic experience. Crucially, bilingual but not monolingual infants discriminated the native vowel contrast, contrasting previous studies in which either delay or equal pace was found for bilingual infants in comparison to their monolingual peers. We hypothesize that the difference may surface with contrasts of low acoustic salience.

Considering that one's native linguistic experience may constrain her speech perception, and bilinguals have more referents to hinge on when perceiving a native contrast, a non-native contrast was examined in Exp.2. Differences between monolingual and bilingual infants were once again observed. Since language-specific perceptual attunement would predict a loss of sensitivity to non-native contrasts, the current differences may be interpreted as a successful perceptual attunement in monolingual infants, but a delayed process in bilinguals triggered by a more complex language environment. Alternatively, infants may have undergone the perceptual attunement of linguistic pitch [24], but bilingual's rich, more varied linguistic experience facilitates their perception of the non-native contrast. In addition, musicianship and learning tone language has been shown to facilitate non-native tone perception in adulthood [25]. The current finding suggests that exposure to two languages may lead to similar outcomes.

In Exp.3, a music contrast was tested with the results showing the similar pattern as in previous two experiments. The null result in monolingual infants suggests that similar to language contrasts, music perception is subject to acoustic salience. It is harder to perceive the inconspicuous violin pitch contrasts. The difference between monolingual and bilingual infants indicates that the bilingual advantage may extend to the musical domain. Furthermore, language and music processing may share common perceptual and neural mechanisms, which may further be affected by the bilingual experience. The detailed separation and overlap between speech and music processing proposed in some previous studies [26,27] need to be explored in future research under a bilingual setting.

From the outcomes of the three experiments, we propose a heightened acoustic sensitivity hypothesis: Bilingual infants may pay more attention to acoustic details in the input than their monolingual peers. This hypothesis may originate from, or be intertwined with: 1) daily experience of a complex language environment; 2) a tightened phonetic space from two languages, forcing bilingual infants to be sharp in detecting native sound patterns; 3) better neural plasticity and less neural commitment, avoiding the formation of false sound categories; etc. Crucially, bilingual infants' heightened acoustic sensitivity is not restricted to the linguistic domain but extends to music perception. The advantage may surface with less salient contrasts as shown in Exps 1 and 2.

The proposed hypothesis may be one of the explanations not only for studies showing bilingual infants' enhanced sensitivity to non-native contrasts compared to monolingual infants [28,29] but also for mixed findings. For initially discriminable contrasts that require realignment or strengthening, too much attention to acoustic details may not help in category formation / boundary stabilization, resulting in (temporary) delay [7,8], and a later sound category formation than monolinguals [30]. Paying more attention to the acoustic cues from the input may be another learning strategy bilingual infants use to keep pace with monolinguals along the developmental trajectory [31].

An alternative hypothesis is that bilingual infants may benefit from their enhanced cognitive abilities, such as executive function and/or attention in the discrimination tasks. Follow-up studies are needed to disentangle these possibilities.

4. Conclusion

Before the first year of life, bilingual infants outperform their monolingual peers in the perception of a native vowel, a non-native pitch, and a musical pitch contrast. The cross-domain perceptual differences between monolingual and bilingual infants are explained by a heightened acoustic sensitivity hypothesis stating that infants growing up in a bilingual environment may pay more attention to acoustic details in the input compared to monolinguals. Previous studies have shown cognitive advantages [11,12] among bilingual infants. The current study extends the advantages to linguistic and musical perception, indicating a cross-domain effect brought by bilingualism in infancy.

5. Acknowledgements

We thank the Babylab research group and the Experimental Phonology group members of UiL-OTS, Utrecht University, SST anonymous reviewers for their valuable comments, and all families that participated in our research.

6. References

[1] Werker, J. F., and Tees, R. C., “Cross-language speech perception: Evidence for perceptual reorganization during the first year of life”, *Infant behavior and development, 7*(1), 49-63, 1984.

[2] Lynch, M. P., and Eilers, R. E., “A study of perceptual development for musical tuning”, *Perception & Psychophysics, 52*(6), 599-608, 1992.

[3] Maurer, D., and Werker, J. F., “Perceptual narrowing during infancy: A comparison of language and faces”, *Developmental Psychobiology, 56*(2), 154-178, 2014.

[4] Hoff, E., Core, C., Place, S., Rumiche, R., Senor, M., and Parra, M. , “Dual language exposure and early bilingual development”, *Journal of child language, 39*(1), 1, 2012.

[5] De Houwer, A., Bornstein, M. H., and Putnick, D. L., “A bilingual–monolingual comparison of young children’s vocabulary size: Evidence from comprehension and production” *Applied Psycholinguistics*, 1-23, 2013.

[6] Liu, L., and Kager, R.W.J., “Are bilingual infants better at learning non-native words contrasted in tones?”, submitted.

[7] Bosch, L., and Sebastián-Gallés, N., “Simultaneous bilingualism and the perception of a language-specific vowel contrast in the first year of life”, *Language and Speech*, 46(2-3), 217-243, 2003a.

[8] Sebastián-Gallés, N., and Bosch, L., “Developmental shift in the discrimination of vowel contrasts in bilingual infants: is the distributional account all there is to it?”, *Developmental science, 12*(6), 874-887, 2009.

[9] Albareda-Castellot, B., Pons, F., and Sebastián-Gallés, N., “The acquisition of phonetic categories in bilingual infants: new data from an anticipatory eye movement paradigm”, *Developmental science, 14*(2), 395-401, 2009.

[10] Sundara, M., and Scutellaro, A., “Rhythmic distance between languages affects the development of speech perception in bilingual infants”, *Journal of Phonetics*, 39(4), 505-513, 2011.

[11] Kovács, Á. M., and Mehler, J., “Flexible learning of multiple speech structures in bilingual infants”, *Science, 325*(5940), 611-612, 2009b.

[12] Singh, L., Fu, C. S., Rahman, A. A., Hameed, W. B., Sanmugam, S., Agarwal, P., ... and Rifkin-Graboi, A, “Back to Basics: A Bilingual Advantage in Infant Visual Habituation”, *Child development, 86*(1), 294-302.2014.

[13] Trehub, S. E., Thorpe, L. A., and Morrongiello, B. A., “ Infants' perception of melodies: Changes in a single tone”, *Infant Behavior and Development*, *8*(2), 213-223, 1985.

[14] Rietveld, T., Kerkhoff, J., and Gussenhoven, C., “Word prosodic structure and vowel duration in Dutch”, *Journal of Phonetics, 32*(3), 349-371, 2004.

[15] Curtin, S., Fennell, C., and Escudero, P., “Weighting of vowel cues explains patterns of word–object associative learning”, *Developmental Science*, 12(5), 725-731, 2009.

[16] Veenker, T., “The ZEP experiment control application”, Utrecht: Utrecht Institute of Linguistics OTS, Utrecht University, 2013.

[17] Liu, L., and Kager, R.W.J., “Perception of tones by infants learning a non-tone language”, *Cognition, 133*(2), 385-394, 2014.

[18] Huang, T., & Johnson, K. (2010). Language specificity in speech perception: Perception of Mandarin tones by native and nonnative listeners. *Phonetica, 67*(4), 243-267.

[19] Kuhl, P.K., Stevens, E., Hayashi, A., Deguchi, T., Kiritani, S., and Iverson, P., “Infants show a facilitation effect for native language phonetic perception between 6 and 12 months”, *Developmental Science*, 9, F13–F21, 2006.

[20] Tsuji, Sho, Reiko Mazuka, Alejandrina Cristia, and Paula Fikkert. "Even at 4 months, a labial is a good enough coronal, but not vice versa." *Cognition 134,* 252-256.

[21] Werker, J.F., and Tees, R.C., “Speech Perception as a Window for Understanding Plasticity and Commitment in Language Systems of the Brain”, *Wiley Periodicals, Inc. Developmental Psychobiology*, 46, 233–251, 2005.

[22] Liu, L., and Kager, R.W.J., “Understanding Phonological Acquisition through Phonetic Perception: The Influence of Exposure and Acoustic Salience”, *Phonological Studies, 18*, 51-58, 2015b.

[23] Best, C. T., McRoberts, G. W., LaFleur, R., and Silver-Isenstadt, J., “Divergent developmental patterns for infants' perception of two nonnative consonant contrasts”, *Infant behavior and development*, 18(3), 339-350, 1995.

[24] Mattock, K., and Burnham, D., “Chinese and English infants' tone perception: evidence for perceptual reorganization”, *Infancy, 10*(3), 24-265, 2006.

[25] Marie, C., Delogu, F., Lampis, G., Belardinelli, M. O., and Besson, M., “Influence of musical expertise on segmental and tonal processing in Mandarin Chinese”, *Journal of Cognitive Neuroscience, 23*(10), 2701-2715, 2011.

[26] Patel, A. D., “Sharing and nonsharing of brain resources for language and music”, From *Language, Music, and the Brain*, edited by Michael A. Arbib, Strüngmann Forum Reports, vol. 10, J. Lupp, series ed. Cambridge, MA: MIT Press, 2013.

[27] Chen, A., Liu, L., and Kager, R.W.J., “Cross-domain correlation in pitch perception, the influence of native language”, *Language, Cognition and Neurosciences*, in press.

[28] Byers-Heinlein, K., and Fennell, C. T., “Perceptual narrowing in the context of increased variation: insights from bilingual infants”, *Developmental psychobiology*, 56(2), 274-291, 2014.

[29] Liu, L. and Kager, R.W.J., “Perception of Tones by Bilingual Infants Learning Non-Tone Languages”, *Bilingualism: Language and Cognition*, in press.

[30] Petitto, L. A., Berens, M. S., Kovelman, I., Dubins, M. H., Jasinska, K., and Shalinsky, M., “The “Perceptual Wedge Hypothesis” as the basis for bilingual babies’ phonetic processing advantage: New insights from fNIRS brain imaging”, *Brain and language*, 121(2), 130-143, 2012.

[31] Mattock, K., Polka, L., Rvachew, S., and Krehm, M., “The first steps in word learning are easier when the shoes fit: Comparing monolingual and bilingual infants*”, Developmental science*, 13(1), 229-243, 2010.

Monolingual and bilingual adults can successfully learn foreign language words implicitly

Hana Zjakic[1,2], *Alba Tuninetti*[1,2], *Paola Escudero*[1,2]

[1]MARCS Institute for Brain, Behaviour, & Development, Western Sydney University, Locked Bag 1797, Penrith, NSW 2751, Australia

[2]Centre of Excellence for the Dynamics of Language, Australian Research Council, Canberra, ACT, Australia

H.Zjakic@westernsydney.edu.au, A.Tuninetti@westernsydney.edu.au, Paola.Escudero@westernsydney.edu.au

Abstract

Assessing word learning methods has important implications for classroom settings and developing language learning programs. Previous research has shown that adults can learn native-language words implicitly [1] and second-language (L2) words explicitly [2]. We tested whether adults could learn L2 words and attend to phonetic detail implicitly using cross-situational word learning, where participants make word-object associations by tracking word-object co-occurrences across learning trials. Results show that participants learned L2 words implicitly with above-chance accuracy, and that the amount of phonetic detail needed to distinguish words determined their accuracy levels. We discuss how these results compare to explicit L2 word learning.

Index Terms: second language learning, word learning, implicit

1. Introduction

Learning words is a challenging process as learners will often be presented with a number of potential referents for a single linguistic label at once [1]. Word learning can occur in two ways: implicit word learning happens without the learner's awareness and learners' access of implicit knowledge is automatic [3]. Alternatively, explicit word learning is a more conscious process and the acquired knowledge can only be accessed when time constraints are not an issue and the learner's focus in on accuracy [4]. Importantly, implicit and explicit word learning are not mutually exclusive as both play a vital role in language learning. However, learners may prefer one method over the other depending on factors such as age, language exposure and experience [5] or instructional context, whether it be formal (e.g. in a classroom) or naturalistic (e.g. in an individual's home) [6].

Previous studies have shown that the acquisition of knowledge becomes more explicit over the course of an individual's lifespan [7]. According to the Maturational Hypothesis [5], children are maturationally constrained to rely on implicit learning mechanisms, while adults rely primarily on explicit methods of learning. For example, Escudero, Broersma and Simon [8] tested native Dutch listeners on their ability to learn novel Dutch words using an explicit word learning task. Participants completed a learning phase in which each trial consisted of a novel image being presented on a screen along with the audio for the corresponding novel word in the form of *'This is an X'*. The same word was then repeated *('Click on the X')*, along with two images appearing on the screen (i.e. the target object and a distractor object) and participants were instructed to select the correct object. During learning, participants were familiarised with all 12 novel words and their visual referents and were then tested on their ability to match the words and objects during a test phase. Results indicated that participants had higher accuracy for the non-minimal pairs (i.e. sets of words that differ in two or more segments: nonMPs) than minimal pairs (i.e. sets of words that differ in only one segment: MPs).

Using a similar task, Elvin, Escudero, Williams, Shaw and Best [2] tested 20 native Australian-English (AusE) monolingual adults on their ability to learn novel Brazilian Portuguese (BP) words. Participants were more accurate at learning nonMPs than MPs. Accuracy rates were also higher for easy-MPs than difficult-MPs, which were categorized according to how difficult the vowel contrasts were to discriminate. The lower accuracy scores for the difficult-MPs provide support for models of speech perception such as the Perceptual Assimilation Model (PAM) [9] and Second-Language Linguistic Perception Model (L2LP) [10, 11, 12] which state that the perception of L2 sounds is influenced by the acoustic properties of an individual's native language (L1). According to both models, when L2 listeners perceive two non-native sounds as one native category (known as single-category assimilation in PAM and NEW scenario in L2LP), they will face the difficult task of either creating a new L2 category for one of the non-native sounds or splitting an existing L1 category. Mapping two non-native sounds on to two separate L1 categories (known as a two-category assimilation in PAM and SIMILAR scenario in L2LP) poses another learning challenge for L2 learners as they must adjust the boundaries of their existing L1 categories to match those of the L2.

Interestingly, previous research has also shown that adults are able to implicitly learn word-object pairings [13] and encode words with fine phonological detail [1] in their L1 via cross-situational word learning (XSWL) – a paradigm which involves listeners inferring word-object associations by tracking word-object co-occurrences across a number of learning trials [14]. Using this paradigm, native AusE monolingual and bilingual adults were tested on their ability to learn novel AusE words [1]. Participants were presented with novel words in CVC structure which formed nonMPs, vowel-MPs or consonant-MPs. Vowel-MPs differed only in their vowel (e.g. DIT-DEET) and consonant-MPs differed only in their initial consonant (e.g. BON-TON). While word learning performance was worse for vowel-MPs than nonMPs and consonant-MPs, listeners performed above chance for all pair types, demonstrating that adults can successfully learn novel L1 words implicitly regardless of their language background.

Therefore, the present study aimed to test whether AusE-speaking monolingual and bilingual adults were able to implicitly learn novel L2 words via XSWL. Using an implicit version of the task reported in Elvin et al. [2], we manipulated the perceptual difficulty of the MPs based on the vowel categories (i.e. easy-MPs, difficult-MPs). The difficult-MPs refer to the vowel contrasts in which both vowels are acoustically closer together based on their F1 (tongue height) and F2 (tongue backness) values, while the easy-MPs refer to the vowel contrasts in which the vowels are acoustically more distinct. We examined participants' accuracy when identifying the novel BP words after exposure to the learning trials. We hypothesized that participants would show an overall accuracy and that both nonMPs and MPs will be above chance, demonstrating that adults can learn foreign language words and at the same time pay attention to phonetic detail when learning words implicitly, that is, without explicit instruction of which word corresponds to which object referent.

In line with Elvin et al.'s [2] findings for explicit L2 word learning, we also predicted higher accuracy for nonMPs compared to MPs, and easy-MPs compared to difficult-MPs. It was further anticipated that there would be no significant difference in the word learning accuracy of the monolinguals compared to the bilinguals [1].

2. Method

2.1. Participants

Participants included 10 first year psychology undergraduates from Western Sydney University, ranging from 17-38 years of age. Five participants were native AusE monolinguals, while five were AusE-speaking bilinguals with intermediate to native-like proficiency in at least one other language. Participants had no prior knowledge of the Portuguese language so as to yield reliable results. Participants recruited through Western Sydney University's research participation system SONA received course credit, while those recruited by word of mouth received monetary compensation for their participation.

2.2. Stimuli

The stimuli for the experiment included 14 novel BP words consisting of the Portuguese vowels /i, e, ɛ, a, o, ɔ, u/ which were used in [2] and were selected from Escudero, Boersma, Rauber and Bion's [15] corpus. Seven target words were in fVfe form and contained each of the seven vowels. The other seven words (kɔko, kuke, pipe, popo, sase, sɛso, teko) were distractor words recorded by a native female speaker of BP. The word pairs comprised 70 nonMPs, 15 easy-MPs and 4 difficult-MPs.

The six difficult vowel contrasts were /a-ɔ/, /a-ɛ/, /e-i/, /o-u/, /e-ɛ/, and /o-ɔ/, based on the close positioning of the vowels with respect to their F1 and F2 values (see Figure 1 which shows the values for the seven Portuguese oral vowels from [15] and the 12 AusE vowels from [16]). The remaining 15 vowel contrasts were considered perceptually easy: /a-e/, /a-o/, /a-i/, /a-u/, /e-o/, /e-ɔ/, /e-u/, /ɛ-i/, /ɛ-o/, /ɛ-ɔ/, /ɛ-u/, /i-o/, /i-ɔ/, /i-u/, and /ɔ-u/, as the vowels in each contrast are acoustically more distinct from each other (see Figure 1). Each word was randomly paired with a novel object. All participants were presented with the same word-object pairings across learning and test trials.

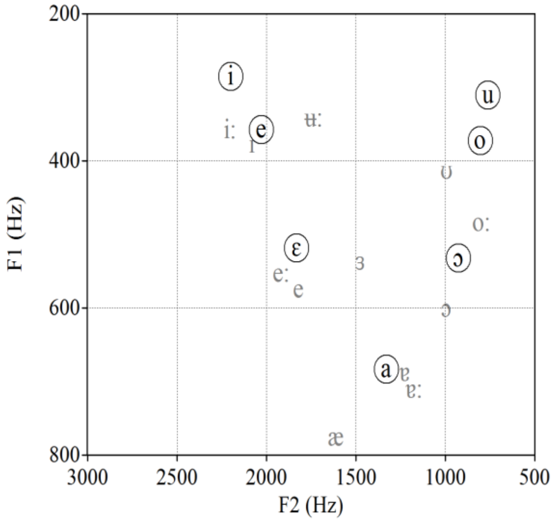


Figure 1: *Male speakers' average F1 and F2 values for BP vowels (black with circles) [15] and AusE vowels (grey) [16]*

2.3. Procedure

The study was an implicit version of the task in [2]. Participants first completed a language background form and provided consent. During learning, participants viewed novel objects and heard novel words on a screen without receiving explicit instruction to make any associations between them. The learning phase consisted of 84 trials, each word appearing as the target six times. During each trial, two images appeared side by side on a white screen (800 x 600 resolution). The audio words for the objects were presented 500 ms after the objects appeared, named in a random order with 500 ms between each word (either right/left or left/right, counterbalanced between participants). Participants were not given any instruction during the learning phase. The word pairs in each learning trial formed nonMPs, easy-MPs or difficult-MPs.

During testing, participants were presented with the same objects and words as those in the learning phase on a laptop. For each trial, two images appeared side by side on the screen. After 500 ms, participants heard an audio word corresponding to one of the objects and had to select whether they thought the word corresponded to the left or the right image via key press. The test phase consisted of 280 nonMP trials (70 pairs x 4 times), 60 easy-MP trials (15 pairs x 4 times) and 24 difficult-MP trials (6 pairs x 4 times), resulting in a total of 364 trials during which each word appeared as a target 26 times. The test phase took approximately 25 minutes for the participants to complete.

3. Results

Participants' accuracy rates were examined in a one sample *t*-test and it was found that participants performed better than chance for all three pair types (nonMPs: *M* = .82, *SE* = .38, *t*[2799] = 44.317, *p* < .001; easy-MPs: *M* = .61, *SE* = .49, *t*[599] = 5.608, *p* < .001 and difficult-MPs: *M* = .57, *SE* = .50, *t*[239] = 2.080, *p* = .04. This illustrates that all listeners were able to infer object-word associations for each pair type. The mean word learning accuracy rates for the monolingual and bilingual groups are illustrated in Figure 2.

The accuracy rates for each language group were analysed in a 2 (Language group: monolingual, bilingual) x 3 (Pair type: nonMPs, easy-MPs, difficult-MPs) repeated measures ANOVA. No main effect of language group was found, *F* (1, 238) = .142, *p* = .707 as monolinguals (*M* = .63, *SE* = .03) and bilinguals (*M* = .64, *SE* = .03) had similar performance in word learning accuracy. A main effect of pair type was found, *F* (2, 476) = 8.748, *p* < .001, such that participants had significantly greater overall accuracy for nonMPs compared to difficult-MPs (β = .171, 95% CI [.072, .270], *p* < .001) and nonMPs compared to easy-MPs (β = .129, 95% CI [.026, .232], *p* = .008). There was no interaction found between pair type and language group, *F* (2, 476) = 1.346, *p* = .261: the mean accuracy scores were similar for all three pair types regardless of language background (monolinguals; nonMPs: *M* = .71, *SE* = .04; easy-MPs: *M* = .64, *SE* = .05; difficult-MPs: *M* = .54, *SE* = .05; and bilinguals; nonMPs: *M* = .77, *SE* = .04; easy-MPs: *M* = .58, *SE* = .05, difficult-MPs: *M* = .59, *SE* = .05.

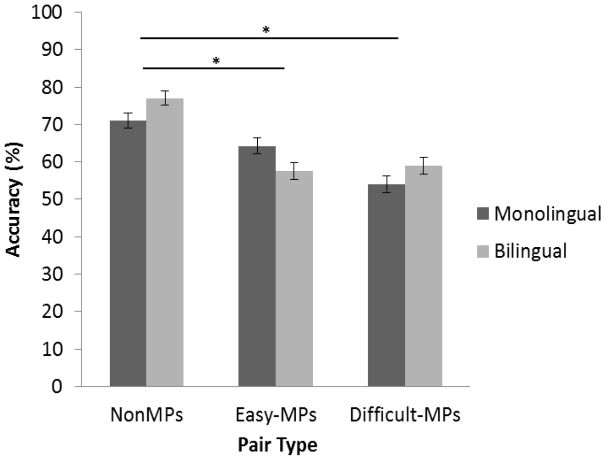


Figure 2: *Mean word learning accuracy rates across the three pair types (nonMPs, easy-MPs, difficult-MPs) for the monolingual and bilingual groups. * represents p < .05*

With respect to the difficult-MPs, no main effect of language group was found, *F* > 1; monolinguals (*M* = .54, *SE* = .05) and bilinguals (*M* = .59, *SE* = .05) had similar performance in accurately learning the difficult vowel contrasts. No main effect of difficult-MP type was found, *F* (5, 190) = 1.339, *p* = .249, and post-hoc pairwise comparisons showed no significant differences between difficult-MP types. There was no interaction found between language group and difficult-MP type, *F* > 1; mean accuracy scores were similar for all six difficult-MP types regardless of language background.

The results show that on average, the bilingual group had difficulty learning the BP difficult vowel contrasts in the following order (from least difficult to most difficult): /e-i/, /a-ɛ/, /a-ɔ/, /e-ɛ/, /o-u/, /o-ɔ/, while the following order was found for the monolingual group: /e-i/, /a-ɛ/, /a-ɔ/, /e-ɛ/, /o-ɔ/, /o-u/.

Table 1: *Average accuracy scores and standard errors for the bilingual and monolingual groups across the six difficult-MP types*

	Difficult-MP Type	Mean (%)	Std. Error
Bilingual	a-ɔ	60	.113
	a-ɛ	65	.109
	e-i	70	.107
	o-u	55	.112
	e-ɛ	60	.113
	o-ɔ	45	.114
Monolingual	a-ɔ	55	.113
	a-ɛ	65	.109
	e-i	65	.107
	o-u	35	.112
	e-ɛ	55	.113
	o-ɔ	50	.114

4. Discussion

The present study shows that adults can successfully learn L2 words through implicit instruction. The finding that listeners were able to learn words regardless of their language background supports the findings of [1] who found that native AusE monolingual and bilingual adults were able to learn novel AusE words using XSWL. While adults rely primarily on explicit learning mechanisms, adults are able to develop explicit knowledge even in implicit training conditions [5] which could explain why the participants successfully learned novel L2 words across all three minimal pair types.

The finding that participants' accuracy scores were lower for the easy-MPs and difficult-MPs across both word learning conditions compared to nonMPs provides support for models of speech perception such as PAM [9] and L2LP [10, 11, 12] which state that L2 learners will initially perceive the sounds in the L2 based on the acoustic properties of their L1 categories. While the BP vowel contrasts in the study are present in English (see Figure 1), they would have been difficult for learners to perceive as the acoustic properties of the vowels vary in each language. The higher accuracy scores for the nonMPs (e.g. fofe-teko) might have been due to the fact that consonants are considered to be more salient in word recognition and have a larger lexical role in English [17]. Therefore, the words that differed in both their consonants and vowel would have been easier to discriminate than words that only differed in their vowel (e.g. fife-fafe). Furthermore, participants' overall lower accuracy on the easy-MPs and difficult-MPs may have been due to the fact that vowels are more important than consonants when identifying words in continuous speech [17] while the present study only presented listeners with isolated novel words.

While it is known that both explicit and implicit methods are viable for word learning [1, 2], there has been little research comparing implicit and explicit word learning in adults. In order to investigate whether adults' word learning performance is better when learning words implicitly or explicitly, we looked at the mean accuracy scores for the 10 participants in the present study and 10 participants from Elvin et al.'s [2] study.

Table 2: *Average accuracy scores for implicit and explicit word learning across nonMPs, easy-MPs and difficult-MPs (2 d.p.)*

	NonMPs	Easy-MPs	Difficult-MPs
Implicit	82.11	61.17	56.67
Explicit	96.21	79.67	60.83

Based on the average word learning accuracy scores for participants in the implicit and explicit conditions, it can be argued that adults perhaps learn L2 words better when receiving explicit instruction rather than implicit instruction. Participants in both conditions were better able to learn words that formed nonMPs than words that formed MPs, and easy-MPs compared to difficult-MPs. While we investigated implicit word learning of individuals with mixed language backgrounds, testing a larger number of AusE monolinguals and then comparing our findings to those of [2] could determine whether one method of word learning is stronger than the other.

While no main effect of BP vowel contrast was found in the present study, [2] found a main effect of BP vowel contrast, suggesting that there are in fact specific vowel contrasts that are easier or harder to learn when compared to the L1 vowel repertoire. Therefore, further testing will allow us to answer whether there are specific BP vowel contrasts that lead to better or worse performance in implicit word learning between monolinguals and bilinguals when the sample size is increased. This will elucidate how specific vowel categories influence word learning. [2] found that on average, AusE listeners had difficulty learning the BP vowel contrasts in the following order (from least difficult to most difficult): /a-ɔ/, /a-ɛ/, /e-ɛ/, /i-e/, /o-ɔ/ and /o-u/ (similar to the current findings). The authors suggested that the vowel contrasts /a-ɔ/, /a-ɛ/ were easier for the listeners to recognise due to the fact that both vowels in the contrasts are acoustically similar to different native categories meaning there is less perceptual overlap with other sounds.

Ongoing research at our lab will also assess infants' ability to learn L2 words via implicit and explicit word learning paradigms and then compare the results to the present findings in order to determine whether there is a developmental trend in learning strategies, as proposed by [5]. Regarding possible limitations of the study, as there were a large number of learning trials (n = 84), some participants may have guessed the purpose of the learning phase, even without receiving explicit instruction to match the words to the objects. Asking participants if they understood the purpose of the learning trials after completion of the task may allow us to examine if there are any differences in performance between participants who had determined the intention of the study and those who had not. Overall, assessing methods of word learning has important implications for classroom settings and the development of language learning programs, and we believe that our results add to the literature on how adults can effectively learn an L2.

5. Acknowledgements

The authors and this work were supported by the ARC Centre of Excellence for the Dynamics of Language [CE140100041]. We would like to thank Jaydene Elvin for her help with the development of the stimuli and data collection.

6. References

[1] Escudero, P., Mulak, K. E., & Vlach, H. A., "Cross-situational learning of minimal word pairs", Cognitive Science, 1-11, 2015.

[2] Elvin, J., Escudero, P., Williams, D., Shaw, J., & Best, C.T., "The role of speech perception and individual differences when learning words in a non-native language: English vs Spanish learners of Portuguese", under review.

[3] Hulstijn, J. H., "Theoretical and empirical issues in the study of implicit and explicit second-language learning: Introduction", Studies in Second Language Acquisition, 27(2): 129-140, 2005.

[4] Ellis, R., "Measuring implicit and explicit knowledge of a second language: a psychometric study", Studies in Second Language Acquisition, 27: 141-172, 2005.

[5] Lichtman, K., "Age and learning environment: Are children implicit second language learners?", Journal of Child Language, 43(3): 707-730, 2016.

[6] Muñoz, C. [Ed.], Age and the rate of foreign language learning (Vol. 19), Multilingual Matters, 2006.

[7] Bialystok, E., "Representation and ways of knowing: three issues in second language acquisition", in N. C. Ellis [Ed.], Implicit and Explicit Learning of Languages, 549-569, San Diego, CA: Academic Press, 1994.

[8] Escudero, P., Broersma, M., & Simon, E., "Learning words in a third language: Effects of vowel inventory and language proficiency", Language and Cognitive Processes, 28(6): 746-761, 2013.

[9] Best, C. T., "A direct realist view of cross-language speech perception", in W. Strange [Ed.], Speech Perception and Linguistic Experience, 171–204, Timonium, MD: York Press, 1995.

[10] Escudero, P., "Linguistic Perception and Second Language Acquisition: Explaining the Attainment of Optimal Phonological Categorisation", Ph.D. thesis, LOT Dissertation Series 113, Utrecht University, Utrecht, 2005.

[11] Escudero, P., "The linguistic perception of similar L2 sounds", in P. Boersma and S. Hamann [Ed.], Phonology in Perception, 152-190, Berlin: Mouton de Gruyter, 2009.

[12] van Leussen, J.-W., & Escudero, P., "Learning to perceive and recognize a second language: the L2LP model revised", Frontiers in Psychology, 6: 1–12, 2015.

[13] Fitneva, S. A., & Christiansen, M. H., "Looking in the wrong direction correlates with more accurate word learning", Cognitive Science, 35(2): 367-380, 2011.

[14] Yu, C., & Smith, L. B., "Rapid word learning under uncertainty via cross-situational statistics", Psychological Science, 18(5): 414-420, 2007.

[15] Escudero, P., Boersma, P., Rauber, A. S., & Bion, R. A., "A cross-dialect acoustic description of vowels: Brazilian and European Portuguese", The Journal of the Acoustical Society of America, 126(3): 1379-1393, 2009.

[16] Elvin, J., Williams, D. & Escudero, P., "Dynamic acoustic properties of monophthongs and diphthongs in Western Sydney Australian English", Journal of the Acoustical Society of America Express Letters, 140(1): 576-581, 2016.

[17] Nespor, M., Peña, M., & Mehler, J., "On the different roles of vowels and consonants in speech processing and language acquisition", Lingue e Linguaggio, 2(2): 203-230, 2003.

The bilingual advantage in the language processing domain: Evidence from the Verbal Fluency Task

Gloria Pino Escobar[1,2], *Marina Kalashnikova*[1], *Paola Escudero*[1,2]

[1]The MARCS Institute for Brain, Behaviour and Development,
Western Sydney University, Australia
[2]ARC Centre of Excellence for the Dynamics of Language, Canberra, Australia

{G.PinoEscobar, M.Kalashnikova,Paola.Escudero}@westernsydney.edu.au

Abstract

Several cognitive advantages in the non-verbal domain have been associated with bilingualism. However, it remains debated whether this advantage also extends to the language processing domain in bilingual children. To assess this, monolingual and bilingual eight-year-old children performed a letter (or phonemic) and a category (or semantic) Verbal Fluency Task (VFT) in order to observe executive functioning under language processing demands. Results showed that bilinguals significantly outperformed monolinguals in both versions of the VFT, demonstrating enhanced lexical processing abilities for bilinguals. These findings will be discussed in view of the bilingual advantage controversy.

Index Terms: bilingualism, children, verbal fluency task, executive functioning, bilingual advantage, attentional control.

1. Introduction

Previous studies claim that bilinguals exhibit more efficient cognitive processes than monolinguals [1, 2]. In fact, abundant research has concluded that bilingual speakers, at specific ages and throughout the lifespan, possess enhanced executive functioning skills when compared to monolinguals [1-3]. Attentional control has been proposed to be the component of executive functioning responsible for this cognitive advantage in bilinguals. This component is the ability to focus on important aspects of a task, overcoming irrelevant distractions [4-6]. Therefore, it has been suggested that bilinguals' constant selection and use of lexical forms from only one of their languages while inhibiting lexical forms of the other, enhances their capacity to process competing responses [1, 2, 6]. This bilingual advantage has been mostly observed using non-language processing tasks, which demand executive functioning in the non-verbal domain [7-11].

However, limited research has been carried out to investigate a possible bilingual advantage in the verbal domain, whereby performance relies not only on participants' attentional control skills but also their lexical competence and lexical retrieval skills. Therefore, the present study investigated whether bilingual children exhibit an enhanced attentional control under lexical processing demands using Verbal Fluency Tasks (VFTs) in two different versions, category (or semantic) and letter (or phonemic) [12-14].

VFTs have been widely used in children [12, 15-17] and adults [13, 18, 19] to measure attention and executive functions that involve lexical processing and retrieval [12, 15]. There are two modalities of the VFT, and each of them activates lexical knowledge, semantic memory, and executive control on different levels [13, 14]. The first modality is category (or semantic), which requires the participant to list words from a semantic category (e.g., clothing items, animals, fruits) in a set period of time (e.g., 1 minute). The executive requirements of VFT-category are similar as the ones for regular speech, and this VFT version is also an effective measure of productive vocabulary size [13, 14]. The second modality is letter (or phonemic). The VFT-letter requires the participant to list words beginning with a letter (e.g., f, m, or p) often imposing restrictions on acceptable responses (e.g., to avoid people and places' names and morphologically related words). The VFT-letter is more effective to measure higher demands of attentional control, along with other executive skills that affect capacity of organisation, monitoring and shifting [13, 14].

Studies in adults using VFT [13, 14] have shown a slight bilingual advantage in the letter version of the task. Luo et al. tested three groups of adults with VFT; i) high vocabulary bilingual group, ii) low vocabulary bilingual group, and iii) a monolingual group [13]. In the VFT-letter, the highly proficient bilingual group performed better than the low vocabulary bilinguals and monolinguals. However, no significant differences were found between groups in the VTF-category. The authors suggest that the letter part of the test incites retrieval interference at a higher processing level, confirming higher cognitive skills in bilinguals [13].

However, recent studies that have used both versions of VTF with bilingual and monolingual children have yielded mixed results [14, 20]. For instance, Kormi-Nouri et al. (2012) tested 1600 Persian monolingual and bilingual children with VTF. Children were recruited from school grades 1 through 4 and were divided in three groups: Turkish–Persian bilinguals, Kurdish–Persian bilinguals, and Persian monolinguals. Results showed a slight bilingual advantage for the Turkish-Persian bilinguals in the letter version of the task. However, the VFT-letter conducted in this study was simplified by not including restrictions on potential answers. In the VTF- category version, monolinguals outperformed both bilingual groups [20]. However, the Kurdish–Persian bilingual group was not as proficient in Persian as the Turkish-Persian group, so that a bilingual advantage could not be found in the former group [14].

The recruitment method adopted and the large number of participants in the study described above may have led to difficulties in controlling for variables such as age, socio-economic status (SES) and level of bilingualism [14], as well as language proficiency in the language tested. Careful control of these variables in experimental designs is essential, as children develop their language and cognitive skills at different time rates. Furthermore, important milestones can occur even in short periods of time. For instance, even at 12 years of age, cognitive processes required for language processing tasks have not yet reached adult performance [12, 16].

With respect to the effect of age, Friesen and colleagues [14] tested bilinguals (of English and another language) and English-speaking monolinguals with the VFT at four different age-groups: seven-year-olds, ten-year-olds, young adults, and older adults [14]. Again, in this study the VFT letter also did not include restrictions for the two younger age groups. It was found that seven-year-old bilingual and monolingual children, with similar English proficiency performed similarly in both versions of the task (i.e., letter and category). In the ten-year-old group, the authors matched low vocabulary (LV) bilinguals with monolinguals because they were not able to collect a sample of high vocabulary ten-year-old bilinguals. Although the ten-year-old LV bilinguals had a lower vocabulary than monolinguals, both groups performed similarly in VTF-letter task. This suggests that LV bilinguals compensated for their smaller vocabulary size with enhanced executive function that aided a more efficient lexical access. Ten-year-old bilinguals also showed effortless word retrieval in comparison with the seven-year-old bilingual group [14]. Finally, in the young adult group that was comprised of high vocabulary (HV) proficient bilinguals and monolinguals, bilinguals performed similarly to monolinguals in VTF-category, but bilinguals outperformed monolinguals in the VFT-letter task. Bilinguals' higher performance was held through to the older adult group; decreasing only slightly with age.

Friesen et al.'s [14] findings suggest that performance in VTF category is related to age and vocabulary knowledge; in contrast, performance in VTF letter is closely related to the degree of bilingualism (i.e., high vs. low proficiency). Their findings also suggest a possible developmental trajectory for the bilingual advantage in the verbal domain, emerging sometime after the age of seven years and before young adulthood.

VFT tasks may yield mixed results in children due to differences in English language proficiency between monolinguals and bilinguals [14]. Indeed, bilingual children's receptive vocabulary in their dominant language is often smaller when compared to monolingual children's vocabulary [18, 19, 21, 22], which poses a challenge for this line of research. However, by the age of eight years, English receptive vocabulary in bilingual children has been reported to reach a similar level as the vocabulary of same-age monolingual peers [22]. Therefore, the present study included eight-year-old children to ensure that differences in performance were not due to differences in English proficiency.

We aimed to investigate whether bilingual children exhibit an enhanced attentional control under lexical processing demands in comparison with their monolingual counterparts. VFT letter and category were used to observe attentional control in the verbal domain. It was expected to find a bilingual advantage in the VFT-Letter because this test requires higher demands on executive functions, which have been observed previously in bilinguals [13, 14]. Contrastingly, it was expected to find similar performances in bilinguals and monolinguals in the VFT-category task because this test requires lower demands of executive functioning, which resemble ordinary speech demands [13, 14].

2. Method

2.1. Participants

Thirty-two children participated in the present study. Sixteen Australian-English (AusE) monolingual children (*M* age = 7 years 10 months, *SD* = 3.5 months) and 16 simultaneous highly proficient bilingual children of AusE and another language (*M* age = 7 years 10 months, *SD* = 3.4 months). Children were recruited from a database of parents who have volunteered to participate in child language research at a university laboratory and through flyers and word of mouth. Groups of monolingual and bilingual children were carefully matched in age, gender (bilinguals: 8 female, 8 male; monolinguals: 8 female, 8 male). The bilingual children were proficient in AusE and one of the following languages: Arabic (5), Spanish (3), Cantonese (2), Mandarin (1), Malay (1), Russian (1), Italian (1), Indonesian (1), and Hindi (1).

2.2. Tasks and procedure

All children completed the VFT task and a test of receptive vocabulary. Two versions of the VFT task were administered: VFT-letter to assess the high-order demands of attentional control and VFT-category to assess the low-order demands of attentional control and lexical access. All testing sessions were conducted in a child-friendly laboratory room. All participants were tested in English by a female Australian English native speaker in order to avoid non-target language intrusion. The experimenter was not involved in the design of the present study or in the recruitment of the participants, and was therefore unaware of each participant's language background.

2.2.1. Vocabulary and language background measures

For the purposes of the present study, the criteria for the selection of the participants included in the bilingual group were:

1. Australian bilingual children of AusE and another language. Participants were eligible if they used the language other than English (in comprehension and production) on an average rate of at least 20% per week;
2. To have been in contact with two different languages from birth (AusE and any other language);
3. To have at least one parent who speaks their heritage language at home on a daily basis;
4. To have a similar proficiency in receptive English vocabulary as their aged-matched monolingual counterparts.

Parents were asked about their children's language exposure to check the eligibility criteria before being invited to take part in the study. Additionally, parents completed a language and family background questionnaire to obtain more detailed information related to language exposure, domains of exposure and parental education, as well as a mean of obtaining further confirmation that participants fulfill the selection criteria.

In order to control for AusE vocabulary proficiency across bilingual and monolingual groups, the standardised Peabody Picture Vocabulary Test, Fourth Edition (PPVT-4) [23] was administered. In this test the child is asked to point or say the number of the picture (out of four pictures) that corresponds to the word spoken by the research assistant. The procedure is repeated in every page of the easel, beginning by the age appropriate set, until reaching eight errors in a set. This test took between 15 and 25 minutes to complete, depending on

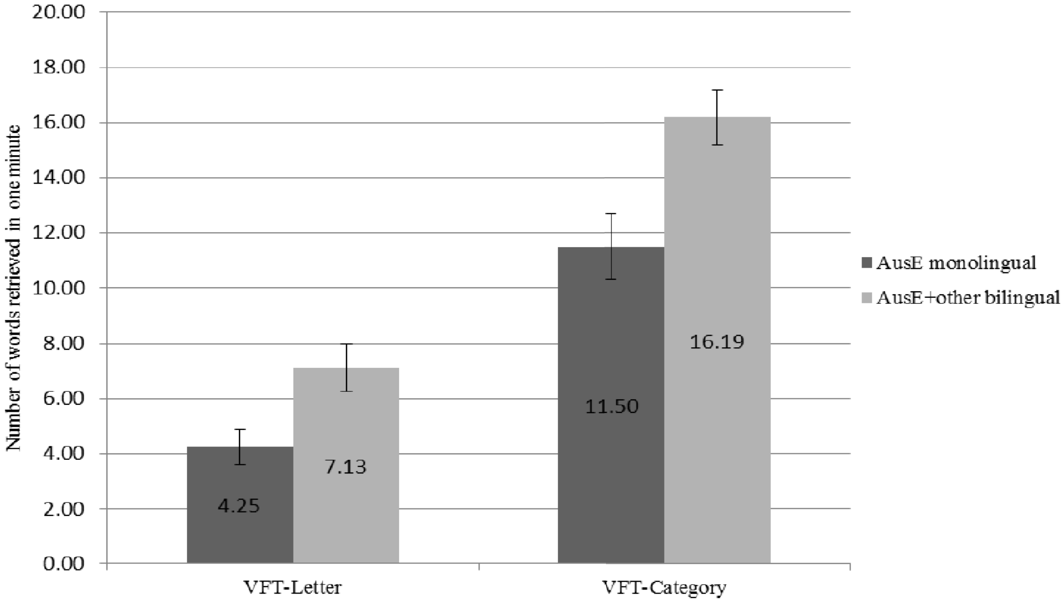


Figure 1: *Number of words retrieved in one minute in the Verbal Fluency letter and category tasks (error bars represent SEM).*

the child's speed and receptive vocabulary proficiency. Standard scores were calculated. No significant differences were found between the bilingual (*M* = 106.38, *SD* = 16.12) and the monolingual groups' (*M* = 107.13, *SD* = 15.47) receptive vocabulary size, *t*(30) = .13, *p* = .89.

2.2.2. *Verbal Fluency Task (VFT)*

In the VFT-category task (or semantic), participants were asked to name all different animals they could think of in one minute. The number of items that the child mentioned in the one minute time frame was counted for the data analysis. Wrong (e.g., words that do not represent an animal) or repetitive answers were excluded from the total score. During the VFT-letter (or phonemic) task, participants were asked to name any word they could think of starting with the letter "f". In addition, they were asked to omit names for people and places, and morphologically related words (e.g., fast, faster, fastest). The number of items that the child mentioned in the one minute time frame was counted for the data analysis.

The order of administration of the tasks was counterbalanced within and across the two language groups (bilinguals: VFT-letter first *n* = 8, VFT-Category first *n* = 8; monolinguals: VFT-letter first *n* = 8, VFT-Category first *n* = 8).

3. Results

The number of correct words retrieved in each VFT task was calculated for the monolingual and bilingual groups, as showed in Figure 1. A repeated measures Analysis of Variance was conducted with VFT condition (VFT-letter, VFT-Category) as the within-subjects factor and language group (monolingual, bilingual) and order of administration of the VFT tasks (VFT-letter first, VFT-Category first) as the two between-subjects factors. Results showed a main effect of VFT condition, *F*(1, 28) = 115.6, *p* < .001, and a main effect of language group *F*(1, 28) = 11.58, *p* = .002. The main effect of task order and all interactions did not reach significance, all *p* > .1. That is, all children obtained significantly higher scores on the VFT category (*M* = 13.84, *SE* = .778) than the VFT letter task (*M* = 5.69, *SE* = .549), and bilingual children *(M* = 11.66, *SE* = .786) outperformed monolinguals (*M* = 7.88, *SE* = .786). In order to further investigate the effect of language group on each task, one-way Analyses of Variance with language group as the independent variable were conducted. The ANOVA for the VFT – letter task showed a significant effect of language group, *F*(1, 31) = 7.168, *p* = .012. Similarly, for the VFT – category task, the ANOVA showed a significant effect of language group, *F*(1, 31) = 9.054, *p* = .005, confirming that bilinguals outperformed their monolingual counterparts in both versions of the VFT.

4. Discussion

The present study aimed to assess whether bilingual children exhibit an enhanced attentional control under lexical processing demands, specifically in the VFT-letter and VFT-category tasks. Our findings showed that bilingual children who were comparable to monolingual children in AusE receptive vocabulary performed better in the two versions of the VFT.

VFT (letter and category) is well known for observing attentional control at different levels [13, 14]; however, VFT also poses a demand for lexical skills, so controlling for vocabulary allowed us to reveal a bilingual advantage in both, the higher and the lower level attentional control conditions. Therefore bilinguals were more efficient than their monolingual counterparts in accessing their lexical repertoire and producing the required words under the different conditions.

Our findings that bilingual children outperformed monolinguals not only in the VFT-letter, but also in the VFT-category task differ significantly from previous studies [14, 20]. However, it is noteworthy that the recruitment criteria in the present study were strict by controlling for English language proficiency, age, and gender across language groups. Additionally, the VFT-letter used for the present study included restrictions such as requiring children to avoid proper nouns and morphologically related words, which made it a more demanding task. Thus, it is possible that this challenging version of the task was more effective in allowing us to detect the differences in performance between the two language groups.

A possible source of these effects is the fact that bilinguals' two languages are permanently activated in the brain requiring them to focus on the target language and disregard the irrelevant language's cues in order to produce speech [1, 3, 5, 6, 24]. Although the VFT is well known for observing attentional control at different levels, it is difficult to disregard other dimensions of executive function involved during this task such as shifting, working memory, and organisation [5, 25-27]. Therefore, it would be interesting to integrate VFT and non-verbal tasks (that are specifically dedicated to testing attentional control) to directly compare the manifestation of the bilingual advantage across the verbal and non-verbal domains.

In summary, it was observed that bilingual children displayed enhanced executive functioning under language processing demands, which indicates a bilingual advantage in the verbal domain. It was also suggested that the bilingual effect on attentional control is more apparent when other variables are controlled. This includes matching for English language proficiency, age, and gender across groups, as well as including highly proficient bilinguals in the bilingual group.

5. Acknowledgements

We thank Aimee Oliveri and Nicole Traynor for their assistance with data collection. We also thank each of our participants and their parents for their willingness to take part in our study. This study was funded by the MARCS Institute for Brain, Behaviour & Development and the ARC Centre of Excellence for the Dynamics of Language.

6. References

[1] Bialystok, E.: 'Bilingualism and the Development of Executive Function: The Role of Attention', Child Development Perspectives, 2015

[2] Kroll, J.F., and Bialystok, E.: 'Understanding the consequences of bilingualism for language processing and cognition', Journal of Cognitive Psychology, 2013, 25, (5), pp. 497-514

[3] Kroll, J.F.: 'On the consequences of bilingualism: We need language and the brain to understand cognition', Bilingualism: Language and Cognition, 2015, 18, (01), pp. 32-34

[4] Bunge, S.A., Dudukovic, N.M., Thomason, M.E., Vaidya, C.J., and Gabrieli, J.D.: 'Immature frontal lobe contributions to cognitive control in children: evidence from fMRI', Neuron, 2002, 33, (2), pp. 301-311

[5] Hernández, M., Martin, C.D., Sebastian-Galles, and Costa, A.: 'Bilingualism beyond language: On the impact of bilingualism on executive control': 'The Cambridge handbook of biolinguistics' (Cambridge University Press, 2013), pp. 161-177

[6] Kalashnikova, M., and Mattock, K.: 'Maturation of executive functioning skills in early sequential bilingualism', International Journal of Bilingual Education and Bilingualism, 2014, 17, (1), pp. 111-123

[7] Bialystok, E., Martin, M.M., and Viswanathan, M.: 'Bilingualism across the lifespan: The rise and fall of inhibitory control', International Journal of Bilingualism, 2005, 9, (1), pp. 103-119

[8] Martin-Rhee, M.M., and Bialystok, E.: 'The development of two types of inhibitory control in monolingual and bilingual children', Bilingualism: Language and Cognition, 2008, 11, (01)

[9] Bialystok, E., and Shapero, D.: 'Ambiguous benefits: The effect of bilingualism on reversing ambiguous figures', Developmental science, 2005, 8, (6), pp. 595-604

[10] Carlson, S.M.: 'Developmentally sensitive measures of executive function in preschool children', Developmental neuropsychology, 2005, 28, (2), pp. 595-616

[11] Bialystok, E., and Senman, L.: 'Executive processes in appearance–reality tasks: the role of inhibition of attention and symbolic representation', Child development, 2004, 75, (2), pp. 562-579

[12] Korkman, M., Kemp, S.L., and Kirk, U.: 'Effects of age on neurocognitive measures of children ages 5 to 12: A cross-sectional study on 800 children from the United States', Developmental neuropsychology, 2001, 20, (1), pp. 331-354

[13] Luo, L., Luk, G., and Bialystok, E.: 'Effect of language proficiency and executive control on verbal fluency performance in bilinguals', Cognition, 2010, 114, (1), pp. 29-41

[14] Friesen, D.C., Luo, L., Luk, G., and Bialystok, E.: 'Proficiency and control in verbal fluency performance across the lifespan for monolinguals and bilinguals', Language, Cognition and Neuroscience, 2015, 30, (3), pp. 238-250

[15] Filippetti, V.A., and Allegri, R.F.: 'Verbal fluency in Spanish-speaking children: analysis model according to task type, clustering, and switching strategies and performance over time', The Clinical neuropsychologist, 2011, 25, (3), pp. 413-436

[16] Matute, E., Rosselli, M., Ardila, A., and Morales, G.: 'Verbal and nonverbal fluency in Spanish-speaking children', Developmental neuropsychology, 2004, 26, (2), pp. 647-660

[17] Ostrosky-Solis, F., Esther Gomez-Perez, M., Matute, E., Rosselli, M., Ardila, A., and Pineda, D.: 'NEUROPSI ATTENTION AND MEMORY: a neuropsychological test battery in Spanish with norms by age and educational level', Applied neuropsychology, 2007, 14, (3), pp. 156-170

[18] Gollan, T.H., Montoya, R.I., and Werner, G.A.: 'Semantic and letter fluency in Spanish-English bilinguals', Neuropsychology, 2002, 16, (4), pp. 562

[19] Sandoval, T.C., Gollan, T.H., Ferreira, V.S., and Salmon, D.P.: 'What causes the bilingual disadvantage in verbal fluency? The dual-task analogy', Bilingualism: Language and Cognition, 2010, 13, (02), pp. 231-252

[20] Kormi-Nouri, R., Moradi, A.-R., Moradi, S., Akbari-Zardkhaneh, S., and Zahedian, H.: 'The effect of bilingualism on letter and category fluency tasks in primary school children: Advantage or disadvantage?', Bilingualism: Language and Cognition, 2012, 15, (02), pp. 351-364

[21] Bialystok, E., Luk, G., Peets, K.F., and Yang, S.: 'Receptive vocabulary differences in monolingual and bilingual children', Bilingualism, 2010, 13, (4), pp. 525-531

[22] Jia, G., Chen, J., Kim, H., Chan, P.S., and Jeung, C.: 'Bilingual lexical skills of school-age children with Chinese and Korean heritage languages in the United States', International Journal of Behavioral Development, 2014, 38, (4), pp. 350-358

[23] Dunn, L.M., Dunn, D.M., and Lenhard, A.: 'Peabody picture vocabulary test: PPVT 4' (Pearson, 2015. 2015)

[24] Hernandez, M., Costa, A., Fuentes, L.J., Vivas, A.B., and Sebastian-Galles, N.: 'The impact of bilingualism on the executive control and orienting networks of attention', Bilingualism: Language and Cognition, 2010, 13, (03), pp. 315-325

[25] Bialystok, E.: 'Bilingualism in development: Language, literacy and cognition' (Cambridge University Press, 2001. 2001)

[26] Hakuta, K., and Diaz, R.M.: 'The relationship between degree of bilingualism and cognitive ability: A critical discussion and some new longitudinal data', Children's language, 1985, 5, pp. 319-344

[27] Hakuta, K., Ferdman, B.M., and Diaz, R.M.: 'Bilingualism and cognitive development: Three perspectives', Advances in applied psycholinguistics, 1987, 2, pp. 284-319

Preliminary performance comparison between PCAKLR and GMM-UBM for computing the strength of speech evidence in forensic voice comparison

Hanie Mehdinezhad [1, 2], *Bernard J. Guillemin* [1, 2]

[1] Forensic and Biometrics Research Group (FaB), University of Auckland, New Zealand
[2] Department of Electrical and Computer Engineering, University of Auckland, New Zealand

Kmah320@aucklanduni.ac.nz,bj.guillemin@auckland.ac.nz

Abstract

A preliminary performance comparison between two probabilistic procedures for the calculation of Likelihood Ratios (LRs) in a Forensic Voice Comparison (FVC) is presented in this paper. One of these, Gaussian Mixture Model–Universal Background Model (GMM-UBM), is common in FVC. The other, Principal Component Analysis Kernel Likelihood Ratio (PCAKLR), is a relatively new procedure. Mel-Frequency Cepstral Coefficients (MFCCs) of three vowels of /aɪ/, /eɪ/ and /i:/ were the speech features used. Scores for each vowel were calibrated and fused using logistic regression. For these experiments PCAKLR is shown to outperform GMM-UBM in terms of both accuracy and reliability.

Index Terms: FVC, MFCCs, GMM-UBM, PCAKLR

1. Introduction

The Bayesian Likelihood Ratio (LR) framework is gaining increased acceptance for evaluating the strength of evidence in a Forensic Voice Comparison (FVC) [1-3]. Multivariate Kernel Density (MVKD) [4] and Gaussian Mixture Model – Universal Background Model (GMM-UBM) [5, 6] are widely used procedures for calculating LRs. The former is primarily designed for token-based analysis, while the latter is primarily designed for data-stream-based analysis [7]. Principal Component Analysis Kernel Likelihood Ratio (PCAKLR) [8, 9], a procedure proposed by researchers at the University of Auckland, is a relatively new approach for computing LRs that is also primarily designed for token-based analysis.

Morrison compared MVKD and GMM-UBM when applied to tokenized data and reported that the later outperformed the former in terms of both accuracy and reliability [10]. Nair et al compared MVKD and PCAKLR for tokenized data and reported that for a large number of input parameters, PCAKLR outperforms MVKD in terms of accuracy [9]. To our knowledge there has not been a similar comparison study between PCAKLR and GMM-UBM. So the goal of this paper is to present a preliminary performance comparison between them for tokenized data.

The remainder of this paper is structured as follows. Section 2 provides an overview of the LR framework, followed by a brief discussion of PCAKLR and GMM-UBM. Section 3 describes our experimental procedure for comparing their performance when applied to tokenized data. The results of these experiments are presented in Section 4, followed by conclusions in Section 5.

2. Background information

2.1. Likelihood Ratio Framework

Mathematically the LR is calculated as:

$$LR = \frac{P(E|H_P)}{P(E|H_d)}.$$

$P(E|H_P)$ is the conditional probability of E (the evidence) given Hp (the prosecution hypothesis) and assesses the similarity between the suspect and offender speech samples. $P(E|H_d)$ is the conditional probability of E given H_d (the defense hypothesis) and measures the typicality of the suspect and offender speech samples to a relevant background population. LR values significantly greater than one support the prosecution hypothesis, LR values significantly less than one support the defense hypothesis, and LR values close to one provide little support either way. It is common to compute the Log-Likelihood-Ratio (LLR), where $LLR = log_{10}(LR)$, its sign indicating whether it supports the prosecution (positive) or defense (negative) and its magnitude indicating the strength of that support.

2.2. Overview of GMM-UBM and PCAKLR

2.2.1. GMM-UBM

GMM-UBM [5, 6] is common in both automatic speaker recognition and FVC. Normally it requires a large amount of data to build a single background model, namely a Universal Background Model (UBM). In order to achieve good performance, the UBM is trained on all background data pooled across speakers. The probability density function of the UBM is estimated using Gaussian Mixture Models (GMMs), with the Expectation Maximization (EM) algorithm being used to train it. The suspect model is then built by copying the UBM and adapting it towards a better fit of the suspect speech data using the Maximum a posterior (MAP) procedure. A score is then calculated as the ratio of the suspect and background probability density function values determined at the offender data points. (Note: A score is calculated using the same expression as for the LR defined above. Once it has been calibrated, it becomes an LR [11, 12])

2.2.2. PCAKLR

PCAKLR is modelled on MVKD [8, 9]. The main difference between the two is that MVKD was designed for a small number of input parameters (typically 3-4), whereas PCAKLR can handle any number of parameters. For both procedures, a normal distribution is used to model the suspect data, while a kernel density distribution is used to model the background data. The distinguishing feature of PCAKLR in

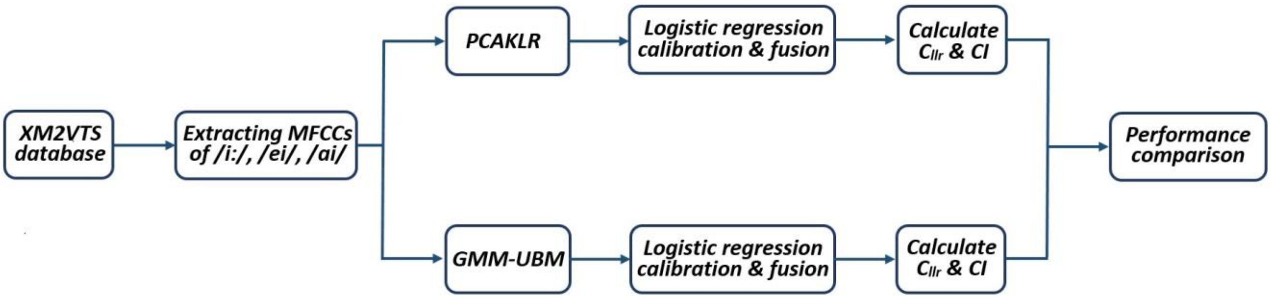

XM2VTS database
Extracting MFCCs of /i:/, /ei/, /ai/
PCAKLR
GMM-UBM
Logistic regression calibration & fusion
Logistic regression calibration & fusion
Calculate C_{llr} & CI
Calculate C_{llr} & CI
Performance comparison

Figure 1: *Experimental set-up for comparing the performance of GMM-UBM and PCAKLR*

comparison to MVKD is the manner in which it takes account of correlations between parameters. It does this by transforming the speech features into a new set of uncorrelated parameters using Principal Component Analysis (PCA). Then individual scores are computed for each of these uncorrelated sets using Univariate Kernel Density (UKD) analysis [13]. Since the transformed parameters are uncorrelated, a final score can be determined by multiplying the individual scores.

2.3. Measuring performance of a FVC / Presenting results

The results of a FVC are often presented using Tippett plots. A Tippett plot, introduced by Meuwly [14], represents the cumulative proportion of the LLR values for both same-speaker and different-speaker comparisons.

The performance of a FVC is measured by evaluating its accuracy and reliability [3]. The accuracy or validity indicates the closeness of the obtained result with the true value of the output. The Log-Likelihood Ratio Cost (C_{llr}) [15, 16] is the recommended metric for assessing this, the lower the value, the better the accuracy. The reliability or precision measures the amount of variation that could be expected in the LR values arising from such factors as the Background set used being necessarily a limited sample of the specified Background population. The Credible Interval (CI) [16] is a popular metric for evaluating this, and again, the lower this value, the better.

3. Experimental Procedure

As shown in Figure 1, using the extracted tokens of /aɪ/, /eɪ/ and /i:/, LRs were calculated using the GMM-UBM and PCAKLR procedures and their performance compared. The following sections expand upon aspects of this process.

3.1. Speech data set

The XM2VTS (Extended Multi Modal Verification for Teleservices and Security) speech database [17] was used in this research. This multi-modal database includes speech recordings digitized at 16 bits and sampled at 32 kHz. The language is English with predominantly a Southern British accent. The database 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 sentences twice. The first two sentences 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 sentence was: “Joe took Father’s green shoe bench out”.

It should be noted that the XM2VTS database contains recordings of read speech and that the level of the background noise is low. Thus from that perspective it is not forensically realistic [18]. However, in support of its use in this investigation, it does consist of a large number of speakers with a similar accent and includes multiple recordings separated by reasonable periods of time, both aspects being important in the FVC arena.

Of the 156 male speakers in this database, only 130 were used in 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 [18] for the rationale behind discarding recordings on the basis of dissimilar accent). Two diphthongs /aɪ/ and /eɪ/ and one monophthong /i:/ were extracted from these recordings from the words “nine”, “eight” and “three”, respectively. So each recording session produced four tokens of each vowel.

Mel-Frequency Cepstral Coefficients (MFCCs) are currently the most popular speech feature used in both the automatic speaker recognition and FVC arenas and have been shown to give good comparison performance [19-21], so they were used in this study also. Their good comparison performance is likely due to the fact that MFCCs are related to the perceptual parameters of the speech signal (i.e., the non-linear response of the human hearing mechanism). We have chosen to use 23 MFCCs in our experiments. This is because the speech data was down-sampled to 8 kHz (a typical value for forensic speech data acquired from landline or mobile phone networks) and at this sampling rate, a maximum of 23 MFCCs can be extracted [22].

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 [11,12].) 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.

Table 1 shows how comparisons were undertaken, the procedure being identical for the Testing and Development sets. In respect to forming the suspect model for each comparison, the data from recording Sessions 3 and 4 were combined, giving eight tokens per vowel. For same-speaker comparisons, Sessions 1 and 2 recordings were used in turn for the offender data. With reference to Table 1, and considering same-speaker

Table 1: *FVC Comparison process*

Speakers	Same-speaker comparisons	Different-speaker comparisons
1	1-S1 vs 1-S34	1-S1 vs 2-S34, 3-S34, …, 43-S34
	1-S2 vs 1-S34	1-S2 vs 2-S34, 3-S34, …, 43-S34
		1-S3 vs 2-S34, 3-S34, …, 43-S34
2	2-S1 vs 2-S34	2-S1 vs 1-S34, 3-S34, …, 43-S34
	2-S2 vs 2-S34	2-S2 vs 1-S34, 3-S34, …, 43-S34
		2-S3 vs 1-S34, 3-S34, …, 43-S34
.	.	.
.	.	.
.	.	.
43	43-S1 vs 43-S34	43-S1 vs 1-S34, 2-S34, …, 42-S34
	43-S2 vs 43-S34	43-S2 vs 1-S34, 2-S34, …, 42-S34
		43-S3 vs 1-S34, 2-S34, …, 42-S34

comparisons for, say, Speaker 43, the two same-speaker comparisons are identified as 43-S1 vs 43-S34 and 43-S2 vs 43-S34. (Note: two same-speaker comparisons are required per speaker in order to compute the CI.) For different-speaker comparisons, Sessions 1, 2 and 3 were used in turn for the offender data. With reference to Table 1, and considering a different-speaker comparison between, say, Speaker 43 (offender) and Speaker 1 (suspect), the three comparisons are identified as 43-S1 vs 1-S34, 43-S2 vs 1-S34 and 43-S3 vs 1-S34. (Again, undertaking multiple different-speaker comparisons for the same pair of speakers is required in order to compute the CI.) With 43 speakers in each of the Testing and Development sets, this resulted in 43 same-speaker comparisons and 903 different-speaker comparisons (ignoring multiple comparisons required in order to compute the CI).

The results for individual vowels were then calibrated and fused using logistic regression [12]. The mean of LRs for the two same-speaker comparisons and the mean of LRs for the three different-speaker comparisons were used to calculate C_{llr}. CI for both same-speaker and different-speaker LRs was computed using the procedure outlined in [16].

Unlike the PCAKLR procedure, the GMM–UBM procedure has two parameters which need to be specified. First is the number of Gaussian components in the GMM. In this investigation this was varied from 8 to 16, the same values as used by Morrison [10]. Second is the number of MAP iterations in the adaptation process. We fixed this at 15, this again being the value Morrison used [10]. (Note: At the outset of our experiments the number of Gaussian components was varied between 1 and 30, and the number of MAP iterations was varied between 1 and 50. The results from these experiments confirmed that the values used in [10] are the best options.) For each vowel, the final choice of number of Gaussian components was made on the basis of lowest resulting C_{llr}, the goal being to try and ensure optimization of the FVC system to this particular data set.

4. Results

The Tippett plots in Figures 2 and 3 show the cumulative distribution of LLR values for GMM-UBM and PCAKLR, respectively. The solid blue curves in these figures are the same-speaker comparison results, and the solid red curves are the different-speaker comparison results. The dashed lines on either side of these solid curves represent the variation in a particular LR comparison result (i.e., LLR±CI). Also shown in these figures are mean C_{llr} and CI.

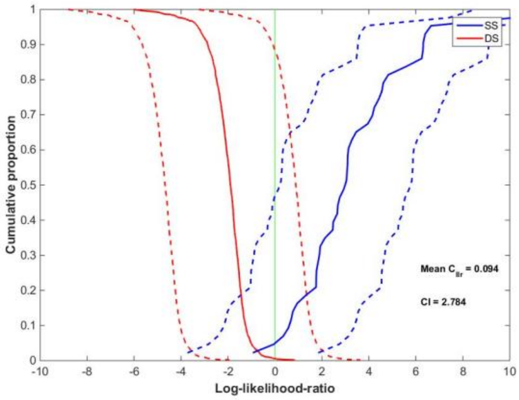


Figure 2: Tippett plot showing the performance of GMM-UBM

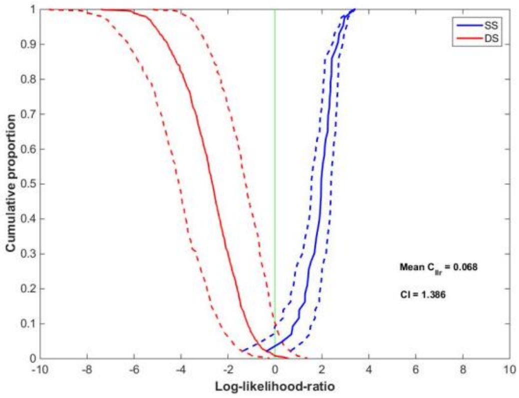


Figure 3: *Tippett plot showing the performance of PCAKLR*

It can be seen from these figures that PCAKLR has marginally outperformed GMM-UBM in terms of accuracy (C_{llr} = 0.068 compared to C_{llr} = 0.094, respectively). This seems to be mainly due to PCAKLR producing a slightly smaller number of same-speaker misclassifications[1], even though the number of different-speaker misclassifications it produced is slightly larger. In terms of reliability, PCAKLR has again outperformed GMM-UBM (CI=1.386 compared to CI=2.784, respectively), the difference now being more significant. The reason for this is not clear and further investigation is needed.

5. Conclusion

A preliminary comparison of the FVC performance of GMM-UBM and PCAKLR when applied to tokenized data has been presented in this paper. The speech feature set used was 23 MFCCs extracted from tokens of the vowels /aɪ/, /eɪ/ and /i:/ spoken by 130 male speakers from the XM2VTS speech database. Results for individual vowels were then calibrated and fused using logistic regression. In terms of both FVC accuracy and reliability, PCAKLR outperformed GMM–UBM, though the improvement in respect to accuracy was only marginal.

6. References

1. Morrison, G.S., *Likelihood-ratio forensic voice comparison using parametric representations of the formant trajectories of diphthongsa).* The Journal of the Acoustical Society of America, 2009. **125**(4): p. 2387-2397.
2. Rose, P. and G. Morrison, *A response to the UK position statement on forensic speaker comparison.* The international journal of speech, language and the law, 2009. **16**(1): p. 139.
3. Morrison, G.S., *Forensic voice comparison and the paradigm shift.* Science & Justice, 2009. **49**(4): p. 298-308.
4. 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.
5. Reynolds, D., *Gaussian mixture models.* Encyclopedia of Biometrics, 2015: p. 827-832.
6. 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.
7. Jessen, M. *Comparing MVKD and GMM–UMB applied to a corpus of formant-measured segmented vowels in German*. in *International Associatio n for Forensic Phonetics and Acoustics Annual Conference (IAFPA 2014), Zurich, Switzerland*. 2014.
8. Nair, B.B., E.A. Alzqhoul, and B.J. Guillemin, *A new approach to computing likelihood ratios based on principal component analysis*, in *UNSW Forensic Speech Science Conference, Sydney, Australia.* . 2012.
9. 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.
10. 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.
11. Enzinger, E., G.S. Morrison, and F. Ochoa, *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 & Justice, 2016. **56**(1): p. 42-57.
12. 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.
13. Lindley, D., *A problem in forensic science.* Biometrika, 1977. **64**(2): p. 207-213.
14. 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.
15. Brümmer, N. and J. du Preez, *Application-independent evaluation of speaker detection.* Computer Speech & Language, 2006. **20**(2): p. 230-275.
16. Morrison, G.S., *Measuring the validity and reliability of forensic likelihood-ratio systems.* Science & Justice, 2011. **51**(3): p. 91-98.
17. Messer, K., et al. *XM2VTSDB: The extended M2VTS database*. in *Second international conference on audio and video-based biometric person authentication*. 1999. Citeseer.
18. Morrison, G.S., F. Ochoa, and T. Thiruvaran. *Database selection for forensic voice comparison*. in *Proceedings of Odyssey*. 2012.
19. Vergin, R., D. O'shaughnessy, and A. Farhat, *Generalized mel frequency cepstral coefficients for large-vocabulary speaker-independent continuous-speech recognition.* Speech and Audio Processing, IEEE Transactions on, 1999. **7**(5): p. 525-532.
20. Ishihara, S. *The Effect of the Within-speaker Sample Size on the Performance of Likelihood Ratio Based Forensic Voice Comparison: Monte Carlo Simulations*. in *Australasian Language Technology Association Workshop 2013*. 2013.
21. Zhang, C., G.S. Morrison, and T. Thiruvaran. *Forensic voice comparison using Chinese/iau*. in *Proceedings of the 17th International Congress of Phonetic Sciences*. 2011.
22. Rabiner, L.R. and B. Gold, *Theory and application of digital signal processing.* Englewood Cliffs, NJ, Prentice-Hall, Inc., 1975. 777 p., 1975. **1**.

Impact of Various GSM Network Factors on Forensic Voice Comparison

Balamurali B. T. Nair [1,2], *Esam A. S. Alzqhoul* [1,2], *Bernard J. Guillemin* [1,2]

[1] Forensic and Biometrics Research Group (FaB), The University of Auckland, New Zealand
[2] Department of Electrical and Computer Engineering, The University of Auckland, New Zealand
bbah005@aucklanduni.ac.nz, ealz002@aucklanduni.ac.nz, bj.guillemin@auckland.ac.nz

Abstract

Speech transmitted through the GSM network can be negatively impacted by a number of factors, such as Dynamic Rate Coding (DRC), Frame Loss (FL) and Background Noise (BN) at the transmitting end. This paper reports on a study to investigate which of these has the greatest impact on the results of a Forensic Voice Comparison (FVC). It is shown that FL tends to have the most significant impact.

Index Terms: Forensic Voice Comparison (FVC), GSM network, Dynamic Rate Coding (DRC), Frame Loss (FL) and Background Noise (BN)

1. Introduction

Mobile phones are now a widely used means of communication amongst the criminal fraternity. In the mobile phone arena there are a number of network systems used worldwide and amongst these the Global System for Mobile Communication (GSM) is currently the most popular with 4.4 billion subscribers [1]. The speech signal transmitted through the GSM network is negatively impacted by a number of factors, such as Dynamic Rate Coding (DRC) [2], Frame loss (FL) [3] and Background Noise (BN) at the transmitting end [4, 5]. Of these, the one that affects FVC the most is unclear and this is investigated in this paper.

To examine the impact of these factors, one can adopt one of two experimental strategies. The first involves transmitting speech through an actual network. This approach, however, has a major drawback in that it permits investigation for only a finite set of transmission conditions existent during a particular call or set of calls. In reality, channel conditions can vary significantly from one call to the next and from one location to the next. For the results of such an investigation to be meaningful, the totality of all possible transmission scenarios needs to be included, not a small subset. Furthermore, with this approach it is not possible to examine the impact of a particular aspect in isolation to others.

The second strategy is to focus on the speech codec implemented in these networks and drive a software implementation of it under all of its possible modes of operation. The rationale for this approach is that while it is the network which dynamically decides the operating mode according to such factors as changing channel quality, it is the codec which implements it. Thus the codec is solely responsible for the quality of the transmitted speech signal [6]. We consider this latter approach to more comprehensively reflect the impact of each and every aspect of the network on speech and for this reason it was chosen for this investigation.

The most widely used speech codec in the GSM network is the Adaptive Multi-Rate (AMR) codec. It operates on speech sampled at 8 kHz and codes it into 20 ms frames. The speech coding technique used is called Algebraic Code Excited Linear Prediction (ACELP) and it can code individual frames at one of eight source coding bit rates: 4.75, 5.15, 5.90, 6.70, 7.40, 7.95, 10.20 and 12.20 kbps [7].

In this investigation the strength of speech evidence was evaluated using the Bayesian likelihood ratio (LR) framework. The LR is a measure of the probability of the evidence (i.e., the suspect and offender data) given the competing same-origin (prosecution) and different-origin (defense) hypotheses [8, 9]. Principal Component Analysis Kernel Likelihood Ratio (PCAKLR) was used to calculate LR values [10].

Mel-Frequency Cepstral Coefficients (MFCCs) are widely used in automatic speaker recognition and FVC for speech acquired from a variety of sources. They have been shown to be optimum when analyzing mobile phone speech as well [11], so were used for this investigation. The accuracy of a FVC analysis has been estimated using the Cost Log-likelihood Ratio (C_{llr}) and its reliability using Credible Interval (CI) [12]. Results have also been shown graphically using Tippett plots [13].

The remainder of this paper is structured as follows. An overview of the factors in a GSM network that can impact a FVC analysis is presented in Section 2. Section 3 discusses the experimental methodology chosen for this investigation. Results and findings are presented in Section 4, followed by conclusions in Section 5.

2. Overview of various GSM factors

2.1. Dynamic rate coding (DRC)

DRC is the process of changing the speech coding bit rate dynamically in accordance with changing channel conditions, and, to a lesser extent, changing channel congestion (i.e., number of users). The coding bit rate in turn directly impacts the quality of the resulting speech signal, and thus any subsequent FVC analysis. With the goal of maintaining a certain minimum speech quality irrespective of changing channel conditions, the GSM network instructs the codec to adjust its coding bit rate either up or down as necessary [2].

DRC is implemented in two stages: channel mode adaptation followed by codec mode adaptation. The first is determined by the number of users (i.e., congestion) in a cell. Full rate (TCH/FR) mode is selected when congestion is low and half rate (TCH/HR) when it is high. With TCH/FR any of the codec's eight bit rates can be used; with TCH/HR this is restricted to the lowest five bit rates. How frequently bit rate can be changed is dependent on which codec mode adaptation procedure is in force. With ETSI-specified fast link adaptation, this is a maximum of every 40 ms, while with Nokia proprietary slow link adaptation the maximum is every 480 ms [14, 15]. A detailed analysis of the impact of DRC alone on FVC can be found in [2].

2.2. Frame loss (FL)

The wireless channel in a mobile network is often very poor, increasing the likelihood of speech frames being either lost or irrecoverably corrupted during transmission. With the GSM network no distinction is made between these two. Whenever a frame is corrupted, an attempt is made to correct it using error correction strategies. If this fails, or the frame is lost, it is synthetically replaced at the receiving end using a history of past ‘good’ frames [16]. When a long sequence of frames is lost, the amplitudes of replaced frames are gradually decreased until silence results or the call is dropped. A detailed analysis of the impact of FL in isolation on FVC can be found in [3]

2.3. Background noise (BN)

BN is frequently present in mobile phone communications. It is different from channel noise in that it originates from a variety of sources at the caller’s location. When added to the speech signal, it negatively impacts the coding process, resulting in a degradation of perceptual speech quality. Though channel noise is also present, it can never directly impact the transmitted speech signal, but rather indirectly by causing frames to get corrupted or lost. As a consequence, all received speech frames are noise-free, though some will be noise-free synthetic replacements for frames either lost or too badly corrupted.

Unlike codecs used in some other networks, the AMR codec has no mechanism for mitigating the impact of BN on the coding process [17]. The severity of this will vary depending on the BN type and the resulting Signal-to-noise ratio (SNR). A detailed analysis of the impact of BN in isolation on FVC can be found in [4].

3. Experiment Setup

3.1. Speech database

The XM2VTS database containing speech recordings of 295 subjects (156 male, 139 female) was used for this investigation [18]. The language is English with predominantly a Southern British accent. Subjects were recorded on four different occasions separated by one month intervals. During each session each subject repeated three “sentences” twice. The first two sentences 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 sentence was: “Joe took Father’s green shoe bench out”.

Of the 156 male speakers, 26 were discarded as either their recordings sounded less audible or they were judged to have a different accent from the rest. The remaining 130 speakers were divided into three sets: 44 speakers in the Background set, and 43 speakers in each of the Testing and Development sets. This resulted in 43 same-speaker (SS) comparisons and 903 different-speaker (DS) comparisons. Further, three different recording sessions permitted two SS and three DS comparisons for every speaker, from which CI values were determined.

Speech files were down sampled to 8 kHz to align with the input requirements of the codec and the three words – “nine”, “eight” and “three” – were extracted from each speaker’s first three recording sessions. Using a combination of auditory and acoustic analysis, the vowels /ai/, /ei/ and /i/ were then extracted. Including diphthongs and a monophthong in the investigation was important because codecs in mobile phone networks often code stationary and non-stationary speech segments differently. A single feature set comprising 23 MFCCs was determined from the entire duration of each vowel segment and used as input to PCAKLR to produce a score. Scores were then calibrated and fused using logistic regression to produce LR values, the required calibration and fusion parameters being determined from the Development set comparisons [19]. A mean LR was calculated for every comparison (i.e., the mean of two LRs for SS comparisons and the mean of three LRs for DS comparisons). These mean values were then used for computing C_{llr}.

3.2. Experimental Methodology

Three FVC experiments were undertaken, as shown in Figure 1. Experiment 1 (top branch) used un-coded speech; Experiments 2 and 3 (middle and lower branches, respectively) used AMR-coded speech. The difference between Experiments 2 and 3 related to which mobile phone factors were incorporated (DRC, FL and BN for Experiment 2; DRC and BN for Experiment 3). In each experiment the Background set was identically processed to the speech being compared.

3.2.1. Implementing DRC

In respect to DRC, a medium-channel-quality scenario has been chosen. With this the codec can use any of its eight bit rates. Which ones it actually uses for a particular call is determined by a complicated process (see [2] for a detailed discussion). But in brief, it involves selecting only four of its eight bit rates, this set being referred to as the Active Codec Set (ACS). With eight bit rates, the total number of possible ACS combinations is 162. One of these combinations was randomly chosen for coding a particular vowel token and it was then dynamically coded using the corresponding four bit rates. The medium-channel-quality scenario also includes a constraint on the initial bit rate to use in a particular call as well as constraints associated with switching to the nearest neighbours in a selected ACS. All these constraints have been included in this investigation.

The ETSI-specified fast link adaptation, permitting bit rate changes a maximum of every 40 ms, has been used. The Nokia proprietary slow LA has not been investigated here as it can be considered to be a subset of the ETSI-specified LA scheme where the chosen codec set contains only one bit rate [2].

3.2.2. Implementing FL

In respect to FL, a frame error rate (FER) in the region of 10 to 15%, which approximately translates into a Mean Opinion Score (MOS) of 2.9 [3], has been used. This equates to the lowest voice quality permitted in such networks. When voice quality drops below this, a call is terminated automatically. Given that the durations of the vowel segments used in these experiments were of the order of 12 to 15 frames, this FER translates into a maximum number of lost frames per vowel segment being typically one, or at most two. Worst-case conditions have been chosen (i.e., two lost frames per vowel token), their locations being determined randomly according to a uniform distribution [3].

3.2.3. Adding BN

In respect to BN, the designers of mobile phone networks typically undertake performance tests using three types of noise: car, babble and street noise, and at three SNRs: 9, 15 and 21 dB [4]. A recent investigation has found that babble

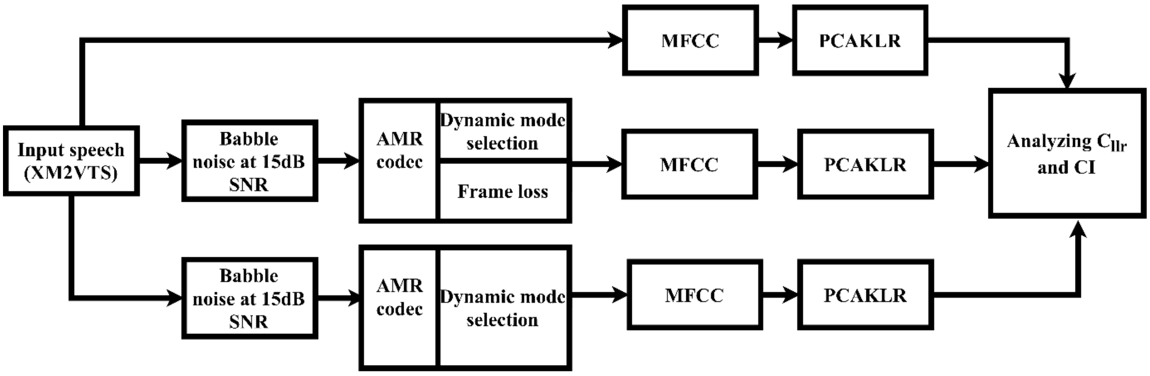


Figure 1: *Block diagram of the experimental setup*

noise tends to have a higher impact on FVC in a GSM network than the other two [4], and so babble noise was used in this investigation. The SNR of the speech signal was set to a medium value of 15 dB.

4. Results

Table 1 shows results of these experiments. Considering first the mean C_{llr} values and comparing performance between un-coded and coded speech, it is evident that AMR coding adversely affects the accuracy of a FVC when DRC, FL and BN are present (compare Experiment 1 with 2 – a lower C_{llr} value translates to better accuracy). However, when DRC and BN are present in the coded speech, but FL is removed, FVC accuracy has very slightly improved compared to the un-coded case (compare Experiment 1 with 3).

Table 1. *Impact of various factors of the GSM network on FVC performance.*

Speech Condition	Mean C_{llr}	CI
Exp. 1: Un-coded	0.167	2.299
Exp. 2: AMR-coded with DRC, FL & BN	0.216	1.627
Exp. 3: AMR coded with DRC & BN (FL excluded)	0.166	1.772

Firstly addressing the slight improvement in FVC accuracy between un-coded and coded speech with FL removed, this is clearly unexpected, but is a result we have observed many times when working with mobile phone speech [2, 3, 4]. We conjecture that it is related to the quantization processes inherent in the AMR-coding which tend to remove small variations in the speech parameters and thereby reduce the differences between voices samples. It seems reasonable to expect this to impact SS comparisons more than DS comparisons, making them even more similar, though both will of course be affected. The fact that, notwithstanding this apparent beneficial impact of the codec, accuracy is quite a bit worse when FL is included in the coded speech, suggests that the impact of FL is more significant than for either DRC or BN.

Focusing now on the CI values in Table 1, it is clear that the reliability of a FVC is better for AMR-coded speech compared to un-coded (the lower the CI value, the better the reliability). We again conjecture that this is due the same quantization processes mentioned above and for the same reasons. But there is little difference in CI for coded speech when FL is present or it is excluded.

Some insight into the above observations can be obtained by examining the Tippett plots for these experiments (Figures 2, 3 and 4 corresponding to Experiments 1, 2 and 3, respectively). These show the cumulative distributions of LLR values for SS (solid blue curve) and DS (solid red curve) comparisons, where $LLR = 10\text{Log}_{10}LR$. (Note: LLRs for correctly classified SS comparisons are positive, those for DS are negative. The greater their magnitude, the stronger the evidence either way). The corresponding dotted lines show the 95% confidence interval for the LLRs.

Comparing first Figure 2 for un-coded speech with Figures 3 and 4 for coded speech, it is clear that the magnitudes of the LLRs for the un-coded case tend to be higher than for the coded, from which it can be concluded that AMR-coding generally negatively impacts the strength of the evidence for a FVC, which is an expected result. Whether this is in part due to the quantization processes mentioned previously is an aspect which deserves further investigation. These plots also confirm that reliability is somewhat better for coded speech than for un-coded.

Comparing now Figures 2 and 3, the proportions of contrary-to-fact LLRs for both SS and DS comparisons are higher for coded speech when FL has been included. Further, SS comparisons have been affected slightly more in this regard than DS comparisons. Comparing Figure 2 with Figure 4 shows that when FL is removed in the coded speech, the proportions of contrary-to-fact LLRs for both coded and un-coded speech are very similar. (Note: to align these observations with the C_{llr} values shown in Table 1, it needs to be remembered that the C_{llr} performance measure gives a greater weighting to LLRs closer to the LLR=0 boundary than those further away, irrespective of whether they are consistent-to-fact or contrary-to-fact.)

5. Conclusions

There are three major factors that impact speech transmitted through a mobile phone network: Dynamic Rate Coding, Frame loss and Background Noise at the transmitting end. This paper has reported on an investigation to determine which of these associated with the GSM mobile phone network has the greatest impact on the performance of a FVC analysis in terms of accuracy and reliability. It has been shown that Frame Loss has the greatest impact on the accuracy of such an analysis and that this seems to be linked to a greater proportion of same-speaker classifications which are contrary-to-fact than different-speaker. In terms of reliability, the coding process associated with this network actually seems to improve this aspect, an observation we conjecture is linked to the associated quantization processes involved.

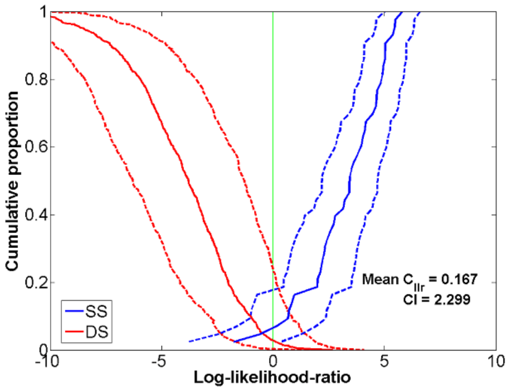


Figure 2: *Tippett plot of the performance for un-coded speech.*

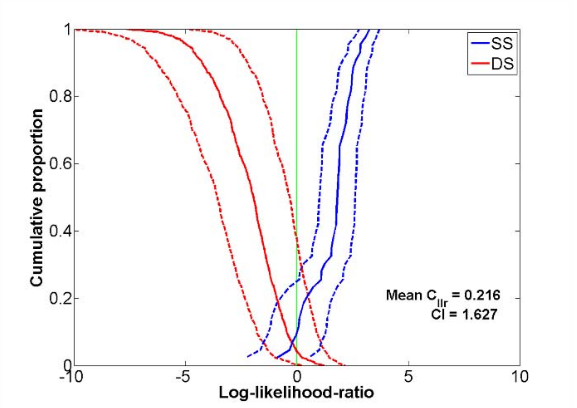


Figure 3: *Tippett plot of FVC performance for coded speech incorporating DRC, FL and BN.*

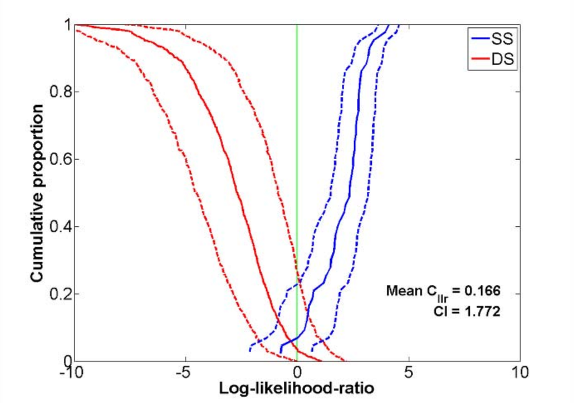


Figure 4: *Tippett plot of FVC performance for coded speech incorporating DRC and BN, but with FL excluded.*

6. References

[1] www.gsacom.com, "GSM celebrates 20 years," *Retrieved on 2 May 2016, last retrieved from http://networks.nokia.com/news-events/press-room/press-releases/gsm-celebrates-20-years,* 2011.

[2] 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.

[3] B. B. Nair, E. A. Alzqhoul, and B. J. Guillemin, "Impact of frame loss aspects of mobile phone networks on forensic voice comparison," *International Journal of Sensor Networks and Data Communications,* Vol. 2015, 2015.

[4] B. B. Nair, E. A. Alzqhoul, B. J. Guillemin "Impact of background noise in mobile phone networks on forensic voice comparison," *J Forensic Leg Investig Sci,* Vol. 2: 007, 2016.

[5] E. A. Alzqhoul, B. B. Nair, and B. J. Guillemin, "Speech handling mechanisms of mobile phone networks and their potential impact on forensic voice Analysis," presented at *Australasian Speech Science and Technology* Conference, 2012, Sydney, Australia, 2012.

[6] E. A. Alzqhoul, B. B. Nair, and B. J. Guillemin, "An alternative approach for investigating the impact of mobile phone technology on speech," in *Proceedings of the World Congress on Engineering and Computer Science, San Francisco, USA*, 2014.

[7] 3GPP, "TS 26.101 V11.0.0 3rd Generation Partnership Project; Technical Specification Group Services and System Aspects; Mandatory speech codec speech processing functions;Adaptive Multi-Rate (AMR) speech codec frame structure. Retrieved on 2 May 2013, last retrieved from http://www.3gpp.org/," ed, 2011.

[8] J. Gonzalez-Rodriguez, P. Rose, D. Ramos, D. T. Toledano, and J. Ortega-Garcia, "Emulating DNA: Rigorous quantification of evidential weight in transparent and testable forensic speaker recognition," *Audio, Speech, and Language Processing, IEEE Transactions on,* Vol. 15, pp. 2104-2115, 2007.

[9] G. S. Morrison, "Forensic voice comparison," *Expert Evidence,* Vol. 40, pp. 1-105, 2010.

[10] 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 & the Law,* Vol. 21, 2014.

[11] E. A. Alzqhoul, B. B. Nair, and B. J. Guillemin, "Comparison between speech parameters for forensic voice comparison using mobile phone speech," *Proceedings of the Australasian Speech Science and Technology Association, Christchurch,* pp. 29-32, 2014.

[12] G. S. Morrison, "Measuring the validity and reliability of forensic likelihood-ratio systems," *Science & Justice,* Vol. 51, pp. 91-98, 2011.

[13] D. Meuwly and A. Drygajlo, "Forensic speaker recognition based on a Bayesian framework and Gaussian Mixture Modelling (GMM)," in *2001: A Speaker Odyssey-The Speaker Recognition Workshop*, 2001.

[14] Nokia, "Guidelines for practical implementation of AMR in Nokia's Network element. General description. Retrieved on 21 June 2013, last retrieved from http://www.scribd.com/doc/104330368/14/Initial-codec-mode-selection," ed, 2004.

[15] 3GPP, "TS 45.009 3rd Generation Partnership Project; Technical Specification Group GSM/EDGE Radio Access Network;Link adaptation. Retrieved on 20 June 2013, last retrieved from http://www.3gpp.org/. ," ed, 2012c.

[16] 3GPP, "TS 26.091 3rd Generation Partnership Project; Technical Specification Group Services and System Aspects; Mandatory Speech Codec speech processing functions; Adaptive Multi-Rate (AMR) speech codec; Error concealment of lost frames. Retrieved on 6 April 2013, last retrieved from http://www.3gpp.org/," ed, 2012b.

[17] 3GPP, "3GPP TS 26.077, Minimum performance requirements for noise suppresser; Application to the Adaptive Multi-Rate (AMR) speech encoder Retrieved on 2 June 2013, last retrieved from http://www.3gpp.org/," 2012.

[18] J. Lüttin, "Speaker verification experiments on the XM2VTS database," in *IDIAP-RR 99-02, IDIAP*, 1999.

[19] D. Ramos-Castro, J. Gonzalez-Rodriguez, and J. Ortega-Garcia, "Likelihood ratio calibration in a transparent and testable forensic speaker recognition framework," in *Speaker and Language Recognition Workshop, 2006. IEEE Odyssey 2006: The*, 2006, pp. 1-8.

Adapted Gaussian Mixture Model in Likelihood Ratio based Forensic Voice Comparison using Long Term Fundamental Frequency

Carolin Elisabeth Buncle Diesner[1], *Shunichi Ishihara*[2]

[1] School of Literature, Languages and Linguistics, the Australian National University, Australia
[2] Department of Linguistics, the Australian National University, Australia
u5820042@anu.edu.au, shunichi.ishihara@anu.edu.au

Abstract

In this paper, the Gaussian Mixture Model – Universal Background Model (GMM-UBM) is applied to one-dimensional speech data, namely the distribution of long term fundamental frequency (LTF0) in likelihood ratio based forensic voice comparison. A series of experiments were conducted using varying numbers of Gaussians, differing adaptation rates to a UBM, and different lengths of speech samples. The results of the GMM-UBM procedure are compared to two previously proposed procedures for LTF0. All three procedures exhibited unique characteristics in their performances. Thus, there was no consistency in performance in that no one procedure constantly outperformed the others.

Index Terms: forensic voice comparison, likelihood ratio, GMM-UBM, long-term F0 distribution

1. Introduction and previous studies

The most common task carried out by forensic phoneticians is forensic voice comparison (FVC) [1]. Different parameters have been used for this task based on, for example, formant frequency, fundamental frequency (F0) contour and the distribution of long-term F0 (LTF0). Formant frequencies have been widely used in FVC, however, when the recording is of poor quality, formant analysis can be difficult. Using formant frequencies can also be rather time consuming. In comparison, [2] suggests F0 as a parameter, as it is generally easy to display and to quantify throughout an utterance. It can be relatively easily extracted even from poor quality recordings, and it is not badly affected by telephone transmission [2]. F0 is also freely available as there are more voiced than unvoiced segments in speech [2]. As a result, the F0-based parameters have been attractive in traditional FVC for some time, although it is important to note that within-speaker variation in F0 can be fairly susceptible to, for example, physiological factors such as age or intoxication, and psychological factors, such as time of day or emotional state [2], [3] and [4].

There are different ways of parameterising the individual unique use of F0 in forensic situations, one of which is modelling F0 contours. This has been proposed in [5] using the *Fujisaki* model. A more commonly used method is looking at distributional patterns of LTF0, the focus of this research. [6] claims that thus far, it was predominantly mean and standard deviation (sd) that have been used to compare the distributional patterns of individual speakers' LTF0. Since this approach (based on mean and sd) seems rather primitive and has provided unsatisfying results, this paper proposed to not only include mean and sd of LTF0 in its likelihood ratio (LR) based FVC approach, but also skewness, kurtosis, modal F0 and probability density of modal F0 (hereafter, the six parameter procedure). All of these parameters relate to the shape of a distribution. The experiments, which used the six parameter procedure with non-contemporary speech samples of 201 Japanese male speakers, succeeded in achieving an overall equal error rate (EER) of 10.7%.

In the above experiments, [6] used running speech between 10 and 25 minutes and did the LR calculations using the multivariate kernel density (MVKD) method [7] in a non-cross-validation manner. Even though this study came up with promising results, the authors criticised the six parameter procedure, as it does not capture the characteristics of bimodal distribution which was mentioned as being very common due to creakiness in voice. Two of the authors went on to do further research and attempted to capture the shape of the LTF0 distribution based on percentiles [8] (hereafter, the percentile procedure). This procedure successfully improved the performance of the system.

Approaches based on the Gaussian Mixture Model (GMM) are commonly used in FVC [9], and, in particular, it was reported that the adapted version of the GMM procedure, namely the Gaussian Mixture Model – Universal Background Model (GMM-UBM) procedure performs well in both automatic [10] and traditional FVC [11].

Despite the fact that the GMM-UBM is almost considered as one of the standard procedures in FVC, to the best of our knowledge, it has never been applied to one-dimensional speech data, such as LTF0. Thus, motivated by [6] and [8], the current study seeks to find out how well the GMM-UBM procedure works with LTF0 within the LR framework. Further experiments using the same database as used in the GMM-UBM procedure, but applied to the procedures proposed in [6] and [8] (the six parameter and percentile procedures), will allow direct comparison of the three tactics.

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) [12].

$$LR = \frac{p(E|H_p)}{p(E|H_d)} \tag{1}$$

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. Experiments

3.1. Database

For the experiments, the monologues of 201 speakers were selected from the Corpus of Spontaneous Japanese (CSJ) [13], which consists of different types of speech samples from 1464 speakers. The selected recordings, the same which have been used in [6], belong to level four or five of the so-called 'spontaneous' scale of either Academic Presentation Speech or Simulated Public Speech. CSJ uses a five-scale evaluation rating of different aspects of the speech used in the recordings. In this case it means that the chosen recordings sound as if they had not been read out, but spoken freely. Using speech samples that involve natural speech rather than speech that is read out, is important as this better simulates forensic casework. Another selection criteria used to determine speech samples was the availability of non-contemporaneous recordings of the same speaker in order to perform within-speaker comparison.

The 201 speakers were further separated into three mutually exclusive databases of test, background and development databases, each of which is made up of 67 speakers. Out of the 67 speakers of the test database, 67 same speaker (SS) comparisons and 4422 independent DS comparisons are possible.

F0 was sampled from each recording at every 0.01 second with the ESPS routine of the Snack Sound Toolkit (http://www.speech.kth.se/snack/). The measured F0 samples were all pooled together for each recording, and they were used to create eight different lengths of samples: 5, 10, 20, 40, 60, 80, 100 and 120 seconds. These different lengths of sample are to investigate the relationship between the performance of a system and the length of samples.

3.2. Likelihood ratio estimation: GMM-UBM

A GMM, which is a parametric probability density function represented as a weighted sum of M component Gaussian densities, is claimed to be the most effective likelihood function in text-independent speaker recognition [10]. GMM parameters (mixture weight, mixture mean and mixture variance/covariance) are estimated from a training database (e.g. suspect samples) using the iterative Expectation-Maximisation (EM) algorithm with the maximum likelihood (ML) estimation. The main idea of the GMM-UBM is that the GMM, which was built by the above process for a suspect, is adapted to a UBM which was built based on the background database. This way of estimating GMM parameters is called Maximum A Posterior (MAP) estimation. Please refer to [10] for a mathematical exposition of the GMM-UBM. In this study, a series of experiments were carried out by altering the number of Gaussian components (2, 3, and 4) and the relevance factor (adaptation weight) (0, 16, 32, 64, and 128). The relevance factor = 0 means no adaptation.

3.3. Likelihood ratio estimation: MVKD

In order to compare the performance of the GMM-UBM procedure with the performances of the procedures proposed in [6] and [8], further experiments were done by replicating [6] and [8] with the same sets of databases. In short, the distribution of the LTF0 was modelled using the mean, sd, skewness, kurtosis, and modal F0 and modal density in [6] (the six parameter procedure), while using the F0 and density values of some percentile points in [8] (the percentile procedure). The MVKD method was used to calculate LRs for both of these procedures. The percentile F0 and its density values were obtained at 5%, 30%, 50%, 70% and 95% for the current study.

3.4. Calibration

A logistic-regression calibration [14] was applied to the derived scores from the three different procedures. The FoCal toolkit (https://sites.google.com/site/nikobrummer/focal) was used for the logistic-regression calibration in this study [14]. The logistic-regression weight was obtained from the development database.

3.5. Performance Assessment

Log-likelihood-ratio cost (C_{llr}) was used to assess the outcomes of the experiments and to gain an overall view of the performance of the systems.

Table 1: *C_{llr} values from the GMM-UBM experiments. Numerals in bold face = lowest C_{llr} value per time unit. Underlined numeral in bold face = lowest C_{llr} value overall*

Gaussians	Adaptation	5	10	20	40	60	80	100	120
2	0	0.881	0.810	0.755	0.710	0.712	0.719	0.705	0.703
2	16	0.868	0.807	0.755	0.711	0.712	0.720	0.705	0.703
2	32	0.860	0.805	0.756	0.712	0.713	0.720	0.706	0.704
2	64	0.853	0.804	0.757	0.713	0.713	0.720	0.706	0.685
2	128	0.852	0.804	0.761	0.715	0.715	0.721	0.707	0.705
3	0	0.873	0.806	0.738	0.714	**0.688**	**0.695**	0.704	0.679
3	16	0.848	0.801	0.745	0.704	0.716	0.719	0.705	0.702
3	32	0.874	0.802	0.756	0.717	0.713	0.710	0.706	0.705
3	64	0.873	0.811	0.761	0.706	0.706	0.724	0.708	0.692
3	128	**0.847**	0.810	0.765	0.723	0.720	0.724	0.710	0.706
4	0	0.854	0.795	**0.733**	0.703	0.709	0.718	**0.670**	**0.667**
4	16	0.873	0.798	0.753	**0.698**	0.708	0.699	0.680	0.678
4	32	0.856	**0.783**	0.749	**0.698**	0.696	0.715	0.684	0.681
4	64	0.895	0.795	0.779	0.707	0.699	0.724	0.689	0.685
4	128	0.889	0.808	0.779	0.714	0.707	0.710	0.718	0.702

◀◀ ◀ Table of Contents for this manuscript ▶ ▶▶

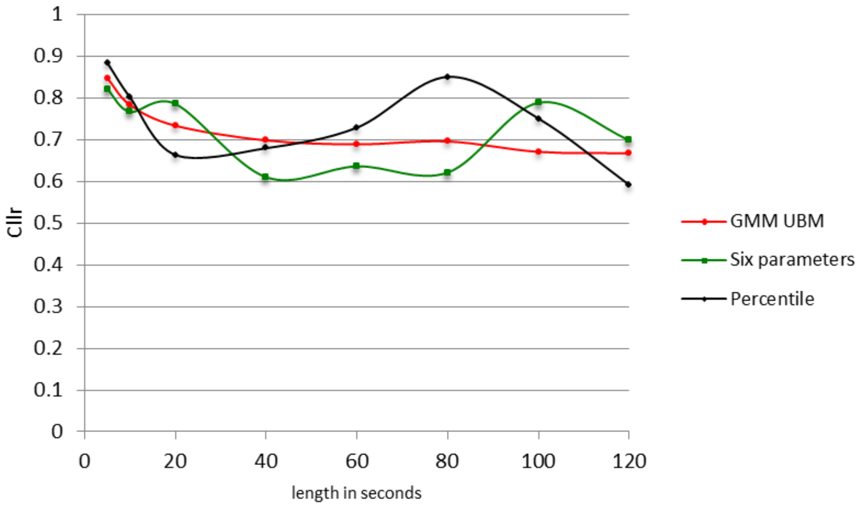


Figure 1: C_{llr} *comparison of all three procedures.*

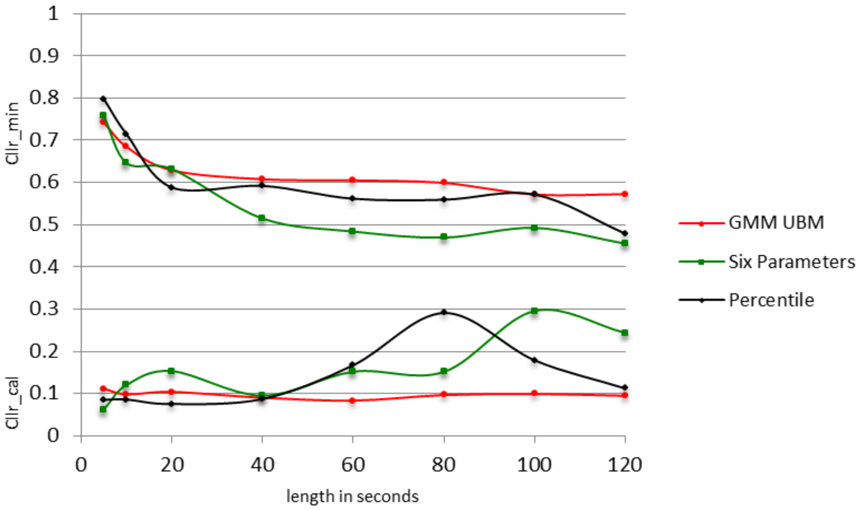


Figure 2: C_{llr_min} *and* C_{llr_cal} *comparison of all three procedures.*

The final C_{llr} value is the sum of two different values: C_{llr_min} and C_{llr_cal}. The former indicates the system's discriminability when it is ideally calibrated and the latter shows the loss caused by its calibration component.

4. Results and Discussion

4.1. GMM-UBM

Table 1 presents the C_{llr} values from the 120 experiments of the GMM-UBM procedure conducted (= 3 Gaussian numbers * 5 relevant factors * 8 sample lengths). Each time unit (0, 5, 10, 20, 40, 60, 80, 100 and 120 seconds) is divided into two, three and four Gaussians with each indicating different adaptation rates starting at zero and rising to 128. The performance improves from 5 to 40 seconds, after which the performance remains relatively stable with some very minor ups and downs in C_{llr}; this can be seen in Table 1 as well as in Figure 1 (red curve), in which the best results (C_{llr} values) per time unit have been plotted. The longer the sample, the more accurately individualising information can be naturally extracted for comparison. Thus, the observation that the performance improves as the sample length becomes longer (in particular from 5 to 40 seconds) is not surprising.

Table 1 further indicates the overall improvement of performance with the increase of Gaussians used per unit of time; the best performance was achieved with the Gaussian number of four, except the sample lengths of 5, 60 and 80 seconds. This was to be expected as an increase in Gaussians increases the chance of better capturing the actual shape of the F0 distribution. Furthermore, it can be observed from Table 1 that within shorter speech samples, a higher amount of adaptation generally improves performance. This comes as no surprise, as the UBM will fill in missing information due to the shorter speech samples. With longer speech samples, a higher amount of adaptation decreases the performance as it then generalises individual information with its adaptation to the UBM. As a whole, the lowest C_{llr} value of 0.667 was achieved with the combination of four Gaussians, zero adaptation and 120 seconds of speech.

4.2. GMM-UBM vs. MVKD based procedures

In Figure 1, the C_{llr} values of the two MVKD-based procedures: the six parameter procedure and the percentile procedure, are plotted against the different lengths of samples. Overall it can be said that these two MVKD-based procedures (green and black curves) also come up with useful results.

Table of Contents for this manuscript

However, they are not as stable as the GMM-UBM (red curve), and the results overall do not necessarily improve as the length of speech samples increases, as one would expect. It can be observed from Figure 1 that the C_{llr} values of the MVKD-based procedures fluctuate a lot from 40 to 120 seconds. Partly because of this instability of the MVKD-based procedures, there is no consistency in performance in that one procedure constantly outperforms the others.

It is interesting to observe that the percentile procedure performed better than the six parameter procedure, only when the sample length is 100 seconds or longer, except for the sample length of 20 seconds. This result contradicts that of [8], which reported that the percentile procedure outperformed the six parameter procedure. This may be due to the fact that [8] used very long sample length (10-25 minutes). Judging from the results of the current study and that of [8], the percentile procedure may work better than the six parameter procedure when the sample is large (100 seconds or longer).

In order to further investigate the instability of the two MVKD-based procedures, the C_{llr} values given in Figure 1 are decomposed into C_{llr_min} and C_{llr_cal}, and they are plotted in Figure 2 against the different lengths of speech samples. If only C_{llr_min} is considered, the three procedures exhibit more or less the same trend; there is a large improvement in performance from 5 to 20 seconds, after which the performance continues to improve overall (with some minor ups and downs) as a function of the sample length, but to a lesser degree. This is something which we expected.

However, unlike C_{llr_min}, it is evident from Figure 2 that the values of C_{llr_cal} fluctuate largely between 0.1 and 0.3 for the two MVKD-based procedures (green and black curves), in particular, with longer samples (60 seconds or longer). It is clear that the instability of C_{llr_cal} is responsible for the irregular pattern observed in the C_{llr} values of the MVKD-based procedures. Yet, the reason behind it is not clear at this stage, and warrants further research.

5. Conclusion and further directions

This paper investigated how well the GMM-UBM is applied to one-dimensional speech data, namely the distribution of LTF0 in LR based FVC. The outcomes of the GMM-UBM procedure were compared to the outcomes of two other, previously proposed procedures: one that models the distribution of LTF0 using the mean, sd, skewness and kurtosis, modal F0 and the density of modal F0 (the six parameter procedure) [6], and the other that models the distribution the F0 and density values of some percentile points (the percentile procedure) [8]. The six parameter procedure and the percentile procedure used the MVKD formula to calculate LRs. The performance was assessed in terms of C_{llr}, including C_{llr_min} and C_{llr_cal}.

As for the performance of the GMM-UBM procedure, it is observed that i) the performance generally improves with the increase of Gaussians used per unit of time, and that ii) with shorter speech samples, a higher amount of adaptation generally improves performance with the combination of four Gaussians.

As for the comparisons between the three procedures, the results (C_{llr}) showed that all three procedures yielded relatively similar results overall. A reason for the similar results could be that they all look at the same thing, that being the distributional pattern of LTF0.

The GMM-UBM seems to be more stable than its competitors, but it does not impress with overall discriminability (C_{llr_min}). The unstable performance of the MVKD-based procedures (the six parameter procedure and the percentile procedure) is due to the inconsistent performance of calibration (C_{llr_cal}) with different lengths of samples. Overall, there is no consistency in performance across the three different procedures; in some cases, one system performed better than the others, in other situations a different system came up with better results depending on the circumstances.

For forensic casework, it is recommended that all three methods be tried to observe how they perform under the particular circumstances of the case.

It is also important to point out that even though the database that was used consisted of highly spontaneous recordings (yet, they are monologues), it would be of high value to do further experiments with more forensically realistic data.

6. Acknowledgements

We would like to thank our anonymous reviewers for their comments. This paper is based on the first author's project report submitted for LING6535 at the ANU.

7. References

[1] P. Foulkes and P. French, "Forensic phonetic speaker comparison," *Oxford Handbook of Language and Law,* pp. 557-572, 2012.

[2] P. Rose, in *Forensic speaker identification*, ed London: Taylor & Francis, 2002, pp. 244-253.

[3] F. Nolan, *The phonetic bases of speaker recognition*. Cambridge: Cambridge University Press, 1983.

[4] A. Braun, *Fundamental frequency - how speaker-specific is it?* Trier: Wissenschaftlicher Verlag, 1995.

[5] A. Leemann, H. Mixdorff, M. O'Reilly, M.-J. Kolly, and V. Dellwo, "Speaker - individuality in Fujisaki model f0 featuers: implications for forensic voice comparison," *The International Journal of Speech, Language and the Law,* vol. 21, pp. 343-370, 2014.

[6] 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.

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

[8] Y. Kinoshita and S. Ishihara, "F0 can tell us more: speaker verification using the long term distribution," presented at the Speech Science and Technology Conference, Melbourne, 2010.

[9] D. Meuwly and A. Drygajlo, "Forensic speaker recognition based on a Bayesian framework and Gaussian Mixture Modelling (GMM)," *Proceedings of 2001 Odyssey-The Speaker Recognition Workshop,* 2001.

[10] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted gaussian mixture models," *Digital Signal Processin,* vol. 10, pp. 19-41, 2000.

[11] G. S. Morrison, "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,* vol. 53, pp. 242-256, Feb 2011.

[12] B. Robertson and G. A. Vignaux, *Interpreting Evidence*. Chichester: Wiley, 1995.

[13] K. Maekawa, H. Koiso, S. Furui, and H. Isahara, "Spontaneous speech corpus of Japanese," presented at the The Second International Conference of Language Resources and Evaluation, Athens.

[14] N. Brümmer and J. du Preez, "Application-independent evaluation of speaker detection," *Computer Speech and Language,* vol. 20, pp. 230-275, Apr-Jul 2006.

Free Labeling of Audio-visual Attitudinal Expressions in German

Hansjörg Mixdorff[1], *Angelika Hönemann*[2], *Albert Rilliard*[3]

[1] Department of Computer Science and Media, Beuth University Berlin, Germany
[2] Faculty of Linguistics & Literary Studies, University of Bielefeld, Germany
[3] LIMSI-CNRS, Orsay, France

mixdorff@bht-berlin.de, ahoenemann@techfak.uni-bielefeld.de, Albert.Rilliard@limsi.fr

Abstract

This paper presents results from a free labeling experiment employing short audio-visual utterances of German produced with varying attitudinal expressions. Raters were asked to freely specify one single word to describe these. Words were classified with respect to emotional dimensions of valence, activation and dominance, as well as assertion/interrogation. As regards modality, video-supported stimuli yielded significantly higher dominance levels than audio-only ones. The main dimensions separating expressions are assertive vs. interrogation, valence, and dominance. The illocutionary strength is associated with the perceived activation, and primarily linked to the visual channel, while sentence mode is primarily conveyed by acoustic cues.

Index Terms: social attitudes, free labeling

1. Introduction

When two talkers converse they always convey information above and beyond pure linguistics, e.g. their mental state, emotions, mood or attitudes. This affective state is influenced, for instance, by the situation or roles of the dialog partners in the social hierarchy. People who share the same language or culture are therefore conditioned to similar codes, behaviors and even belief systems. In contrast, interaction between partners from different cultures may lead to wrong interpretations of social expressions. A study investigated twelve social attitudes e.g. surprise, irritation, command-authority for prosodic effects in the languages British English, French and Japanese [1]. They found similarities across these languages, but also some culture-specific uses of prosodic parameters. The similarities may be explained within the framework of a theory such as the frequency code [2] which proposes the use of pitch level as a marker inverse to dominance. Other codes have been proposed [3] that may refine the predicted use of pitch for communicative purposes. Conversely, culture-specific uses have been documented [4]. Intercultural comparison of linguistic and paralinguistic effects has enjoyed growing attention as the knowledge about how verbal and non-verbal social affects are expressed in different languages is paramount for mutual understanding between different cultures.

The current work is based on the framework developed by [5] in which attitudes are characterized by a situational description of between whom and where they occur. Recordings also concern the visual channel, as facial gestures are known to be a vital part of attitudinal expressions [6].

Attitudes such as arrogance, politeness, doubt or irritation - see Table 1 for abbreviations henceforth used in this paper - were elicited through short dialogs which ended in the target sentences ‘Eine Banane’ (engl. *a banana*) or ‘Marie tanzte’ (engl. *Marie was dancing*). Preceding the target dialog a test dialog was performed in order to prepare the speakers and help them immerse themselves in the context of the attitude.

In earlier perception studies we had native German subjects rate the credibility of the expressions portrayed by the first 10 of the speakers [7]. We then examined the acoustic-prosodic properties of the data and determined the respective differences between types of attitudes [8]. Finally, we ran an identification study in which we asked subjects to choose from a set of five labels the one they deemed most appropriate [9].

Both latter studies showed that attitudes essentially cluster in several groups, the members of which share similar properties. On the positive side of the spectrum we find attitudes such as *admiration* and *sincerity*, whereas *authority*, *contempt*, *arrogance*, *irritation* and to a certain degree *irony* gather on the negative side. “Neutral” statements and questions which we initially regarded as a standard are often confounded with their affective partners *politeness* and *surprise*, respectively. Due to the experiences with the identification study we suspected that offering raters a set sub-group of labels introduces a strong bias. Therefore we decided to follow the approach by [10]. In this study, raters were free to select a single word, either a noun or adjective that best fit their impression of the attitudinal expression. Different from [10], we also included audio-only and video-only examples to test for differences in the modalities. Our methodology for evaluating the results is also slightly different.

2. Perception Study

We selected the stimuli for the study based on our earlier results regarding the performance of the speakers. Eventually stimuli of the best 15 speakers were included. Of these we chose those examples that had been rated best for a given attitude, yielding 6 stimuli for each attitude. In our previous work we found that there was no significant difference of the target utterances in the judgment of the raters therefore we decided to use only the utterance ‘Eine Banane’ to reduce the amount of stimuli. A sub-set of the selected stimuli was added in audio-only and video-only mode. In total we had 96 audio-visual (AV), 48 audio-only (AU) and 48 video-only (VI) samples which we split into two sets of 80 stimuli each.

Further developing the design adopted in [10], a presentation software was developed that included audio-visual, audio-only and video-only stimuli. A warm-up phase was added in which eight stimuli were displayed to familiarize subjects with the range of expressions they were going to rate, however, without asking their assessment. The ultimate task was to describe each of the stimuli with a single word, either a noun or adjective. As mentioned earlier, every subject had to rate 80 examples (48 AV, 16 AU, 16 VI) for the experiment. Warm-up stimuli were presented only in the audio-visual modality and not used in the experiment proper. The rating procedure was allowed to take as long as the subject required. It took between 25 and 45 minutes to complete the task. Subjects were students (30 male, 5 female) of Media Informatics in their second year at the Department of Computer Science and Media at Beuth University Berlin. Participants received course credits in exchange for their time.

3. Normalization and Semantic Analysis of Labels

We collected a total number of 2732 labels: 1631 for AV, 546 for AU and 595 for VI presented stimuli. Analysis of written expressions showed quite a variation of terms used, yielding a total number of 647 different tokens. Despite the instruction to use just a single word, some subjects had entered two or even a whole phrase to describe their impressions. Oftentimes two-word terms included an emotional and a linguistic component, such as "genervt fragend" (engl. *asking irritably*). After the correction of typos we normalized the raters' inputs by collapsing similar words, for instance, such as "Frage" (*question*) and "fragend" (*asking*) onto a single term. We also collapsed semantically equivalent terms onto the more frequent one, e.g. "akzeptierend" (*accepting*), "bestätigend" (*confirming)* and "zustimmend" (*approving*) were collapsed to the more frequent term "zustimmend". The term "fragend" was the most frequently chosen term (N=292), followed by "genervt" (*irritated*, N=149) and überzeugt (*convinced*, N=115) of the 127 non-neutrally perceived terms. Following are the top three expressions across the attitudes depending on the modality: AV: fragend (questioning), N=171 genervt (irritated), N=103, zweifelnd (doubting), N=74, AU: fragend, N=67, erstaunt (surprised), N=34, gelangweilt (bored), N=27, VI: fragend, N=54, entschlossen (determined), N=27, genervt, N=27. There were a small percentage of terms (2.43 %) that we were unable to interpret sensibly and hence failed to map onto any of the normalized expressions. These were all single-occurrence tokens that we excluded from further analysis. After consolidating all expressions we yielded 117 terms.

In order to further pull apart the linguistic and affective content of each normalized term and become more independent of the respective word identity for the ensuing analysis, we classified them following the scheme developed by [11][12][13]. In principle, we analyzed each term in the three-dimensional space of valence, activation and dominance and added to these the linguistic dimension of statement vs. interrogation. We restricted this classification to three possible values: negative, neutral and positive for valence and – , 0 and + for activation and dominance. Such a semantic classification [O] permits the analysis of the emotional and linguistic weight of each term with respect to its frequency of occurrence for a given rater and attitude without being tied to the original term. For a term such as "genervt" (*irritated*), for instance, we assigned negative valence, +activation and +dominance. Depending on the stimulus used to elicit the term in the given case we assigned the sentence mode, here statement. In order to calculate the position assumed by each attitude in the three-dimensional emotional space we mapped the three values onto a scale from -1 to +1 and averaged over all ratings for that given attitude.

4. Results of Analysis

Based on the frequency and semantic values of labels assigned to each attitude we yielded centers of gravity in the emotional space for each attitude. Table 1 lists the positions of all 16 attitudes in the emotional space for audio-visual stimuli. We can see, for instance, that CONT is judged more negatively than AUTH, while POLI has an almost neutral connotation. Based on these results we also compared the impact of reduced modalities on the assessment of attitudes. The result presented in

Figure *1* only concern the subset of utterances presented audio-visually, audio-only as well as video-only.

Table 1: Sixteen attitudes and respective abbreviations, Positions of sixteen attitudes in the emotional space.

attitude	abbrev-iation	valence	activat-ion	domin-ance
admiration	ADMI	.5347	.7030	-.0198
arrogance	ARRO	-.2885	.4423	.4615
authority	AUTH	-.4078	.3398	.4369
contempt	CONT	-.6700	.6100	.6000
neutral statement	DECL	-.1456	.0194	.0000
doubt	DOUB	-.3462	.5096	-.3173
irony	IRON	.1053	.6737	.0842
irritation	IRRI	-.7767	.7961	.6893
obviousness	OBVI	-.3529	.5294	.3431
politeness	POLI	.0577	.1250	.1923
neutral question	QUES	-.1471	.0294	-.0294
seductiveness	SEDU	.6600	.6000	.1600
sincerity	SINC	.3564	.4455	.2277
surprise	SURP	-.0385	.4904	-.1731
uncertainty	UNCE	-.3725	.0686	-.2157
walking-on-eggs	WOEG	-.6117	-.0097	-.1553

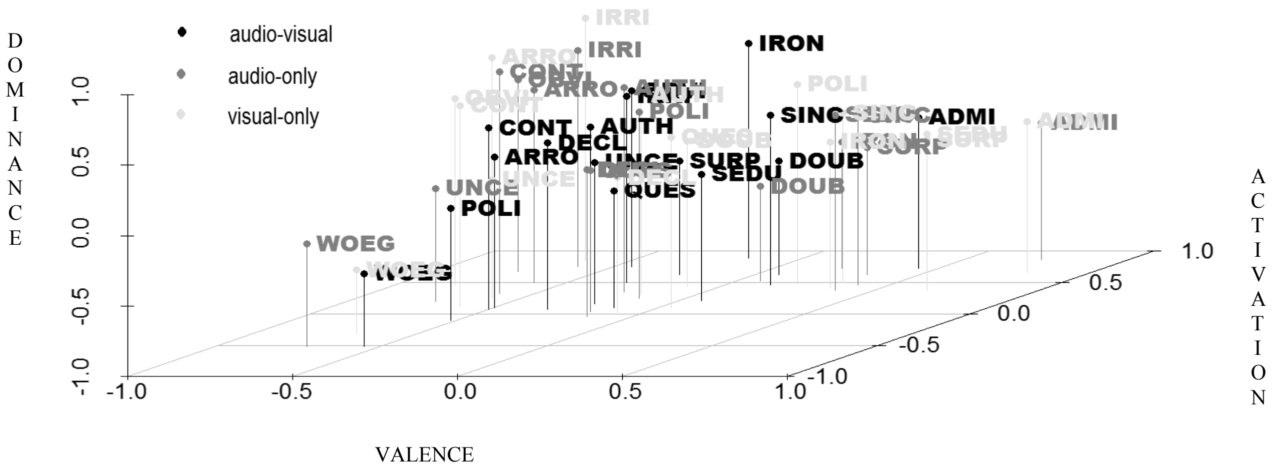


Figure 1: Position of attitudes in the emotional space, subset presented audio-visually, audio-only and video-only.

We will discuss the differences between modalities in detail later in this paper.

Normalized labels were organized in a contingency table with the presented stimuli's 16 categories, in each of the three presentation modalities in rows (i.e. rows present the expressive behavior of speakers, ordered by presentation modality), and the labels assigned in columns (i.e. columns present what was perceived from the stimuli). An analysis of the distribution of these expressive behaviors according to the labels was performed using a correspondence analysis (CA) [14]. The CA was run on the results obtained on the audio-visual modalities only with audio-only and video-only results used as supplementary individuals. The semantic classification of labels according to their valence, dominance, activation and linguistic mode were used as supplementary variables. An elbow criterion indicates to keep the first four dimensions (which explain 56% of the variance) of the CA for further analysis,. Table 2 reports the coordinates and quality of representation ($\cos^2$) of supplementary semantic labels attributed to each labels. This allows us to interpret the abstract dimensions without referring directly to the respective labels collected. The first dimension is mostly linked to the linguistic distinction between assertive and interrogative terms and to the dominance dimension. The second dimension relates to valence, while the third is linked to +activated labels. The fourth is related to expression with -activation.

The first dimension of the CA is mainly built on the expressions of AUTH, CONT, IRRI (on the assertive and +dominant side) and of DOUB, SURP, UNCE and the interrogative and –dominant side. Dimension 2 contrasts SEDU and ADMI (labeled with positive valence labels) to the others. The third dimension – linked to +activated labels – separates IRRI and CONT from the more neutrally perceived expressions of POLI and DECL. The fourth sets apart SEDU and WOEG - being -activated expressions - from OBVI, IRON and IRRI as non-minus activated expressions (but not necessarily +activated).

Table 2: Coordinates and $\cos^2$ ($\cos^2$ are multiplied by 100 and rounded for convenience) of the supplementary semantic categories of labels, on the first 4 dimensions of the CA.

coord.		**Dim. 1**	**Dim. 2**	**Dim. 3**	**Dim. 4**
Val.	neg	0.2	-0.3	0.3	0.0
	neu	0.0	-0.2	-0.4	0.1
	pos	-0.3	0.7	0.1	-0.1
Activ.	-	-0.1	-0.3	-0.4	0.4
	0	0.0	0.0	-0.4	0.0
	+	0.0	0.1	0.4	-0.1
Dom.	-	-0.5	-0.4	-0.1	0.1
	0	-0.2	0.2	-0.2	0.0
	+	0.6	-0.1	0.4	0.0
Ling.	Ass.	0.4	0.3	0.0	0.0
	Int.	-1.0	-0.8	0.0	0.1

$\cos^2$		**Dim. 1**	**Dim. 2**	**Dim. 3**	**Dim. 4**
Val.	neg	9	**43**	24	0
	neu	1	10	**60**	2
	pos	12	**80**	1	2
Activ.	-	2	14	15	**16**
	0	1	1	**64**	0
	+	0	7	**79**	5
Dom.	-	**42**	26	1	1
	0	24	25	28	1
	+	**62**	3	28	0
Ling.	Ass.	**55**	33	0	1
	Int.	**55**	33	0	1

In order to reach a better representation of the multi-dimensional spread of expressions, we applied a hierarchical clustering on the distribution of rows obtained with the first four dimensions of the CA (cf. [14]). Results of this clustering summarize the observed spread of expressions in a 7-cluster solution. In the following we list these clusters with the most frequent labels (in decreasing order of importance, English translations given in italics), semantic connotation, and primarily associated attitudes and their modalities:

Cluster #1: fragend (*asking*), zweifelnd (*doubting*), erstaunt (*astounded*), überrascht (*surprised*), unwissend (*unknowing*); interrogation, –dominant ; DOUB (AU, AV, VI), QUES (AU, AV), SURP (AU, AV), UNCE (AU, AV, VI)

Cluster #2: fragend, unsicher (*unsure*), zurückhaltend (*restrained*), ängstlich (*afraid*), verachtend (*contemptuous*), enttäuscht (*disappointed*); interrogation, -activation, -dominance and 0 valence; WOEG (AU, AV, VI)

Cluster #3: amüsiert (*amused*), erfreut (*pleased*), ironisch (*ironic*), fröhlich (*cheerful*), begeistert (*zealous*), schwärmend

(*enthusiastic*), erzählend (*narrating*), erleichtert (*relieved*), verwirrt (*confused*); positive valence, +activation, assertion, and 0 dominance; SEDU (AU, AV)
Cluster #4: erregt (*aroused*), verführerisch (*seductive*), geheimnisvoll (*mysterious*), begeistert (*excited*), sinnlich (*sensual*), amüsiert, fröhlich, freundlich (friendly); positive valence, 0 dominance, assertion and +activation; ADMI (AU, AV, VI), IRON (AU, AV, VI), SEDU (VI), SURP (VI)
Cluster #5: arrogant (*arrogant*), überzeugt (*convinced*), offensichtlich (*obvious*), zustimmend (affirmative), abfällig (*condescending*), erschrocken (*scared*); assertion,+dominance; ARRO (AV, VI)
Cluster #6: neutral, feststellend (*ascertaining*), gelangweilt (*bored*), bestimmend (*determining*), berichtend (*reporting*);
0 activation, neutral valence, assertion, and 0 dominance; CONT (AV, VI), IRRI (AV, VI), OBVI (AU, AV, VI)
Cluster #7: genervt (*irritated*), aggressiv (*aggressive*), wütend (*furious*), verärgert (*angry*), entschlossen (*determined*), fordernd (*demanding*), autoritär (*authoritarian*); +dominance, negative valence, +activation, assertion; ARRO (AU), AUTH (AU, AV, VI), CONT (AU), DECL (AU, AV, VI), IRRI (AU), POLI (AU, AV, VI), QUES (VI), SINC (AU, AV, VI)

5. Discussion and Conclusions

Audio-only presentations of IRRI, ARRO and CONT (semantic impositions of the speaker on the interlocutor) did not pertain to the same cluster as their audio-visual counterparts. Video-only presentations of SEDU, SURP and QUES, and both mono-modal presentations of IRON are also judged differently than their respective audio-visual versions. Audio-only ARRO and CONT are judged less dominant than audio-visual presentations; audio-only performances of IRRI are perceived as less negative. In these three cases a reduction of perceived illocutionary strength is observed in the AU modality as compared to VI and AV presentations. Video-only SURP and QUES are not understood as interrogations – thus the absence of the acoustic channel makes it more difficult to decode the linguistic meaning (difficult, but not impossible, as video-only DOUB is classified correctly). VI-only SEDU conveys an interrogative meaning, but lacks the joyful and sexually-oriented aspects that are conveyed when the audio modality is present.

The description of IRON by the listeners is interesting, as it is a complex construct: it is described by [15] as a meaning contrasted by prosodic means. For English there does not seem to be one single reliable strategy to express irony. Rather, prosody is used to create a contrast where IRON is to be conveyed. It seems the contrast lies in a mismatch between modalities: Ironic single-modality performances are perceived as dominant (i.e. expressing one type of dominant expression that, however, is not ironic), whereas their combination conveys a positive valence. Irony emerges from this contrast between the interpretations of the two modalities, like a contradiction between modalities (cf. also [16]). Another noticeable fact from the interpretation of irony by German listeners is in contrast with similar tests in French and German [12][17]: the French linked it to obviousness, thus lacking the positive character observed here, and the Japanese rated irony solely as negative, something to be avoided in a conversation. This positive evaluation of irony also contrasts with the results obtained in the categorical perception test, where the "ironic" label was mixed with more negative labels, while the same performance are judged here positively: this bias between perceptual protocols may illustrate a difference between tests that focus on predefined concepts (categorical recognition) vs. tests that allow judging prosodic performances (like this free labeling one).

In contrast to [9], attitudes WOEG, ARRO and SEDU now occupy separate clusters indicating an unambiguous assignment of the labels. In contrast, in our previous experiment no attitude formed its own cluster and e.g. the positive attitude SEDU was mixed with negative attitudes such as ARRO, AUTH and CONT. SEDU and WOEG yielded very low recognition scores. In the previous work SURP and DOUB built one cluster but the same cluster also included POLI, OBVI and CONT which was not plausible. As can be seen in the current clustering SURP, DOUB are joined with the other interrogative attitudes UNCE and QUES. This suggests that free descriptions of attitudes yield more plausible classifications.

6. References

[1] Shochi. T.. Rilliard. A.. Aubergé. V. & Erickson. D. "Intercultural perception of English. French and Japanese social affective prosody". in S. Hancil (ed.). The Role of Prosody in Affective Speech. Linguistic Insights 97. Bern: Peter Lang. AG. Bern. 31-59. 2009.

[2] Ohala. J. J.. "The frequency codes underlies the sound symbolic use of voice pitch". in Hinton. L.. Nichols. J. & Ohala. J. J. (eds.). Sound symbolism. Cambridge University Press. Cambridge. 325-347. 1994.

[3] Gussenhoven. C., *The Phonology of Tone and Intonation*, Cambridge: Cambridge University Press. 2004.

[4] Léon, P., "*Précis de Phonostylistique. Parole et Expressivité,* Paris: Nathan Université, 1993.

[5] Rilliard, A., Erickson, D., Shochi, T., de Moraes, J.A., "Social face to face communication - American English attitudinal prosody", INTERSPEECH 2013. 1648-1652.

[6] Swerts, M. and Krahmer, E., "Audiovisual prosody and feeling of knowing", Journal of Memory and Language 53(1): 81-94, 2005.

[7] Hönemann, A., Mixdorff, H., Rilliard, A. "Social attitudes - recordings and evaluation of an audio-visual corpus in German", Forum Acusticum 2014, Krakow, Poland.

[8] Mixdorff, H., Hönemann, A., Rilliard, A., "Acoustic-prosodic Analysis of Attitudinal Expressions in German." Proceedings of Interspeech 2015, Dresden, Germany, 2015.

[9] Hönemann, A., Rilliard, A., Mixdorff, H., "Classification of Auditory-Visual Attitudes in German." FAAVSP 2015, Vienna, Austria, 2015.

[10] Guerry, M., Shochi, T., Rilliard, A., and Erickson, D. "Perception of prosodic social attitudes affects in French: A free-labeling study Proceedings of ICPhS 2015, Glasgow, Scotland.

[11] http://tschroeder.eu/weblog/?page_id=2, accessed on 5 February 2016.

[12] Schauenburg, G., Ambrasat, J., Schröder, T., von Scheve, C., & Conrad, M. (2015). Emotional connotations of words related to authority and community. *Behavior Research Methods, 47*, 720-735.

[13] Schröder, T., Hoey, J., & Rogers, K. B. (in print). Modeling dynamic identities and uncertainty in social interaction: Bayesian affect control theory. *American Sociological Review*.

[14] Husson, F., Lê, S., Pages, J., "Exploratory multivariate analysis by example using R". London: Chapman & Hall, 2011.

[15] Bryant, G. A., "Prosodic contrasts in ironic speech," Discourse Processes, 47(7), 545-566, 2010

[16] González-Fuente, S., Escandell-Vidal, V. & Prieto, P., "Gestural codas pave the way to the understanding of verbal irony," *Journal of Pragmatics*, 90, 26-47, 2015.

[17] M. Guerry, A. Rilliard, D. Erickson, T. Shochi, "Perception of prosodic social affects in Japanese: a free-labeling study", In Proc. Speech Prosody 2016, Boston, 811-815, 2016.

An Analysis of Lombard Effect on Thai Lexical Tones: The Role of Communicative Aspect

Chariya Boontham[1], *Chutamanee Onsuwan*[1,2], *Tanawan Saimai*[3], *Charturong Tantibundhit*[2,3]
[1]Department of English and Linguistics, Faculty of Liberal Arts, Thammasat University, Thailand
[2]Center of Excellence in Intelligent Informatics, Speech and Language Technology and Service Innovation (CILS), Thammasat University, Thailand
[3]Department of Electrical and Computer Engineering, Faculty of Engineering, Thammasat University, Thailand
babblenoise@gmail.com, consuwan@tu.ac.th, sync_nero@hotmail.com, tchartur@engr.tu.ac.th

Abstract

This work investigates Lombard Effect (clean versus babble) on the realization of Thai lexical tones across two settings: with conversation partner (map task) and without partner (sentence reading). For both, thirty location names whose final syllable varying in lexical tones were constructed and used. Ten pairs of Thai adults participated in the study. The findings showed that in noise condition, regardless of whether the tones were produced with or without conversation partner, F0 values for all tones were significantly higher. Importantly, Lombard Effect on the tones was significantly increased for all but low tone when a partner was involved.

Index Terms: Lombard Effect, Thai, babble noise, lexical tone, map task

1. Background

In 1911, Lombard conducted a study and made an important discovery that during conversation with noise (of various types), speakers adjust their use of language and speech patterns [1]. From then on, the concept of Lombard Effect/Lombard speech was originated, which was the phenomenon where a speaker increases loudness and modifies her speech in noise condition to allow a listener to understand the intended message [2]. Lombard Effect has received extensive attention from speech scientists and relevant studies have been carried out in many languages (e.g., English [3], Spanish [4], Cantonese [5]).

By and large, it was clear that apart from an increase in intensity, speakers make other significant speech adjustments in noise condition (as opposed to clean condition). Specifically, it was found that speakers made adjustment in their production of vowels in terms of vowel duration and intensity [6]. Fundamental frequency and formant frequency were also higher when communication was changed from clean to noise condition [6]. Importantly, these effects were reported across languages (e.g., Spanish [4], French [7]).

However, several studies concluded that Lombard Effect did not take place as a consequence of noises only. They asserted that thoughts, mental states, and reactions of conversation partners also had influences on speech modifications [8]. Lane and Tranel [9] reported that in identical noise conditions, speakers with larger number of conversation partners made greater degrees of speech adjustment_than those in ones-sided communication. In addition, Fitzpatrick, Kim and Davis [10] concluded that during communication in noise condition with conversation partners, vowel duration and F0 were significantly higher than without partner. Those previous findings highlighted an important aspect of communication in Lombard speech where there is at least a 2-way communication and the sender aims to make the receiver understand what she wants to say. Thus, the sender is willing and ready to adapt her speaking strategies for the effectiveness of communication [1]. Moreover, Zeng and Liu [11] stated that individual's speaking patterns/forms usually change abruptly when noises occur during communication or when there are errors in speaking and listening from conversation partners.

As mentioned, many acoustic characteristics were reported to associate with Lombard speech. Recently, Lombard Effects on lexical tones have been explored in tonal languages, such as Cantonese and Thai. Zhao and Jurafsky [5] compared speech adjustments of 6 lexical tones in clean versus white noise conditions (with no conversation partner) from 8 Cantonese speakers aged between 20-52 years old. The findings showed that in white noise condition, F0 in all tones increased, especially in mid and mid-rising tones.

Kasisopa et al. [12] examined speech production in noise, focusing on lexical tones in Thai. Speech data was obtained from six females aged between 27-34 years old in clean and white noise conditions. It was found that F0 and tone contours for the five tones in isolated words were higher than those in sentence frame. Importantly, it was found that F0 for four tones (mid, low, high, and rising tones) in white noise condition was higher than in clean condition.

Lombard Effect not only reflects speech adjustment but also reveals aspects of human speech perception process. A number of studies have investigated perception patterns of Thai lexical tones in noise conditions (e.g., white noise in Onsuwan et al. [13] and pink noise in Mixdorff et al. [14]). To our knowledge, only the study by Kasisopa et al. [12] has addressed the question of Lombard Effect in the production of Thai lexical tones. Therefore, we would like to explore this further by 1) using various types of noise: clean, white, and babble (results in babble noise are only presented here) 2) including two communicative settings (with and without conversation partner). To give readers a background of Thai lexical tones, F0 contours of the five tones (mid, low, falling, high, and rising) from this present study in clean condition where there was no conversation partner are given in Figure 1.

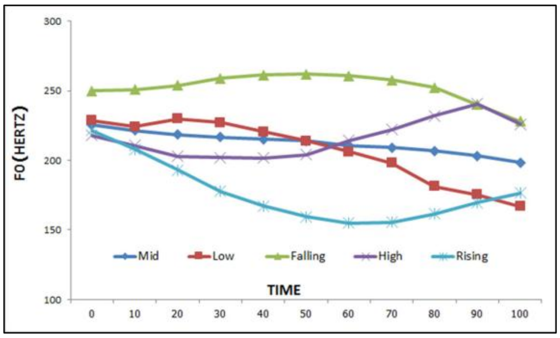


Figure 1: *F0 of five Thai lexical tones (mid, low, falling, high, and rising) from this study in clean condition with no conversation partner.*

2. Method

2.1. Participants

Participants consisted of 10 conversation pairs (one member is a 'speaker' and the other 'listener') aged between 20-40 years old; all pairs had at least one female member (male-female or female-female). Members in each pair know each other well and only one female member was assigned the speaker role.

2.2. Speech materials

Ten target syllables were /pa:/ /pà:/ /pâ:/ /pá:/ /pǎ:/ /na:/ /nà:/ /nâ:/ /ná:/ /nǎ:/. Thirty location names were constructed, each with a target syllable in a final (stressed) syllable. For example, /khrua.khun.**ná:**/ 'Auntie's kitchen', /hâ:ŋ.tà.wan.**ná:**/ 'Tawanna shopping mall'. A same list of target location names was used in map task and sentence reading.

2.3. Data recording

Data were collected from two separate settings: with conversation partner (map task) and without partner (sentence reading). Each recording session, which includes map task and sentence reading, always started with map task and lasted about 20 to 30 minutes.

2.3.1. Map task (with conversation partner)

Following a map task designed by Viethen, Dale and Cox,F [15], three sets of maps were developed; each set consisted of two corresponding maps, one for the speaker and the other the listener as shown in Figures 2 and 3. Three different sets of maps (3 scenes: city scene, beach scene and country scene) were created so that three different noise conditions (clean, white noise, and babble noise) could be randomly be distributed across 10 participant pairs. Before the task begun, each participant was assigned a role, either as a speaker or a listener. They were informed that their task was for the speaker to guide the listener from a starting point (only indicated on the speaker's map) to a finishing point (only indicated on the speaker's map) via a specified route (only indicated on the speaker's map) for each map set. Each map set has 18 assigned locations (10 of which were target location names and 8 were distractors). They were also told that one of them would be hearing some noise from headphones. One short practice trial with a simplified set of maps was given to each pair.

The members were in two separate rooms and could only communicate through microphones and headphones. Only one member (the speaker) heard the noise from headphones. Three types of noise condition were introduced: clean, white noise, and (Thai) multi-talker babble noise (for this paper, only results from babble noise were reported) at 60-75 dBSPL. Recordings of speech from both members were made, but only those from the speakers were analyzed. It should be noted that for each map set, when multiple repetitions of target location names were elicited, but only the best two for each were selected.

2.3.2. Sentence reading (without conversation partner)

The speakers were asked to read at a normal speed 72 sentences (with 30 target location names and 24 distractors) for three times in each noise condition (randomly assigned). The sentence frame was /t͡ɕ^{h}ǎn.paj … ʔ a:t^h ít.ní:/ 'I go to…this week.'

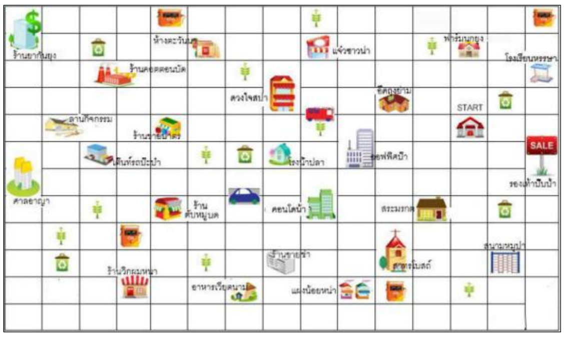


Figure 2: *Sample of listener's map (city scene).*

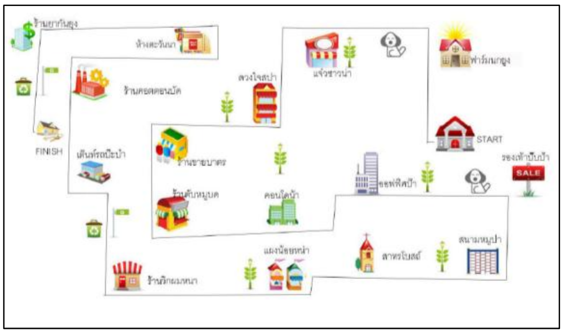


Figure 3: *Sample of speaker's map (city scene).*

2.4. Data analysis

Target syllables were manually segmented and two parameters were measured, i.e., segmental duration (not reported here) and F0 using Praat [16]. Specifically, the F0 values of tone contour of each syllable were extracted in ten equidistant (time-normalized points) from 0% to 100% as shown in Fig. 4. In total, speech tokens analyzed here are composed of 400 tokens from map task (10 syllables × 2 repetitions × 2 noise conditions × 10 speakers) and 600 tokens from sentence reading (10 syllables × 3 repetitions × 2 noise conditions × 10 speakers).

Then, the data were separately analyzed in five repeated three-way ANOVA, i.e., one for each of the five Thai tones. Three independent variables in each of these analyses are Condition (Clean/Lombard), Communication (Map/Sentence), and Time points (10% points), respectively.

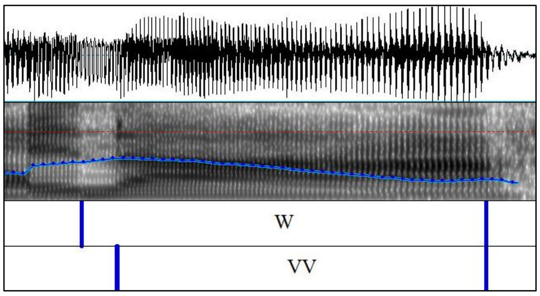


Figure 4: *Waveform and spectrographic display (with F0 line) of a syllable /nàː/ produced in clean condition in sentence reading.*

3. Results

Table 1 shows significant F-values for all factors and their interactions for each of the five tones on fundamental frequency (F0) by comparing the clean and noise condition and with (map) and without (sentence) conversation partners. Findings are explained below.

Clean vs. Lombard: The results revealed that in the babble noise condition, F0 for all of the five tones were significantly higher than those in the clean condition, both in map and in sentence settings (see Table 1). This can clearly be seen when comparing between the two solid curves (map Lombard vs map clean) and between the two dashed curves (sentence Lombard vs sentence clean) in Figures 5-9.

Map vs. Sentence: For map task (both in clean and noise), F0 for all but falling tones were significantly higher than those in sentence setting. The Clean/Lombard × Map/Sentence interactions were significantly different for four tones, i.e., mid, falling, high, and rising, showing that Lombard Effect was stronger in map than in sentence settings. (see Table 1). As for F0 shapes, when comparing the two solid curves (map) with the two dashed curves (sentence), the dashed ones appear to be flatter and lower, but overall shapes remain relatively the same for mid, falling, and high tones (Figures 5, 7 and 8). A notable pattern can be seen in rising tone where the second half of the F0 contours showed a gradual fall rather than a rise in sentence setting.

Trend Analyses: The trend analyses showed that there were significant trends over time for all but high tone. Interestingly, trends interacted with Map/Sentence for mid, falling, and rising tones.

Table 1: *Significant F-values for all factors and their interactions for each of the 5 tones.*

	Mid	Low	Falling	High	Rising
C vs L	**6.94**	**9.02**	***34.58***	***52.20***	***47.89***
M vs S	***61.46***	***18.02***	N/S	***30.11***	***57.43***
TP	*7.74*	***23.07***	***11.54***	N/S	***23.73***
C/L x M/S	***14.03***	N/S	*6.39*	**10.55**	***11.89***
C/L x TP	N/S	N/S	N/S	N/S	N/S
M/S x TP	**2.42**	N/S	***3.62***	N/S	***20.09***
C/L x M/S x TP	N/S	N/S	N/S	N/S	N/S

Note: C stands for speech in clean condition; L stands for Lombard speech; M stands for map task; S stands for sentence reading; TP stands for time points; and N/S stands for not significant. *Italic figure* = *p*<.05; **Bold figure** = *p*<.01; and ***Bold Italic figure*** = *p*<.001.

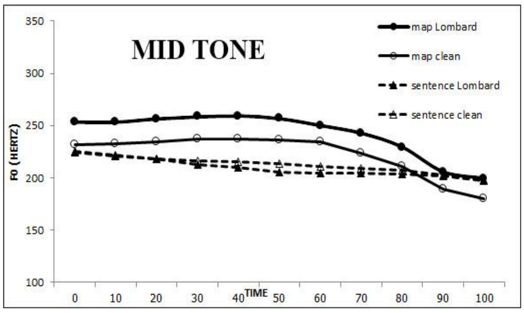


Figure 5: *F0 curves of mid tone across 10 normalized time points in clean vs. babble noise (Lombard) and with conversational partner (map) vs. with no partner (sentence).*

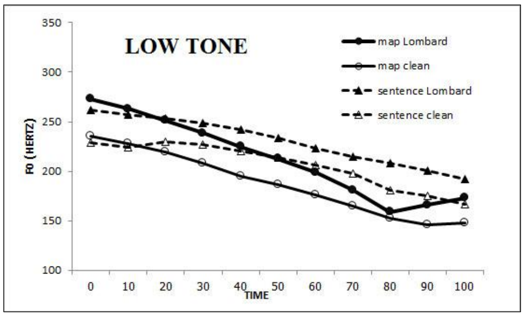


Figure 6: *F0 curves of low tone across 10 normalized time points in clean vs. babble noise (Lombard)and with conversational partner (map) vs. with no partner*

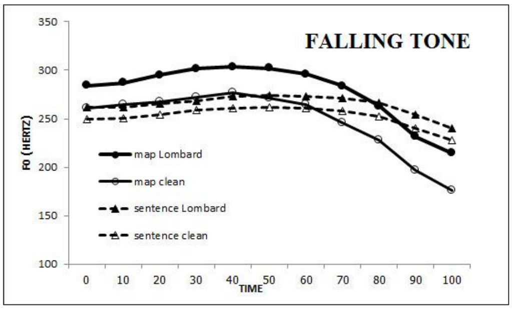


Figure 7: *F0 curves of falling tone across 10 normalized time points in clean vs. babble noise (Lombard) and with conversational partner (map) vs. with no partner (sentence).*

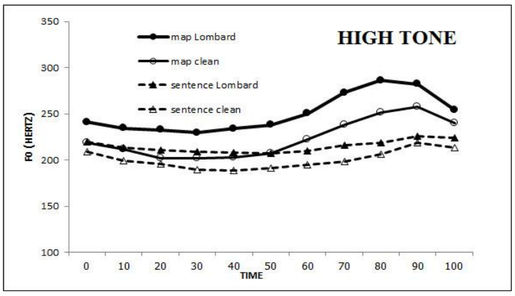


Figure 8: *F0 curves of high tone across 10 normalized time points in clean vs. babble noise (Lombard) and with conversational partner (map) vs. with no partner (sentence).*

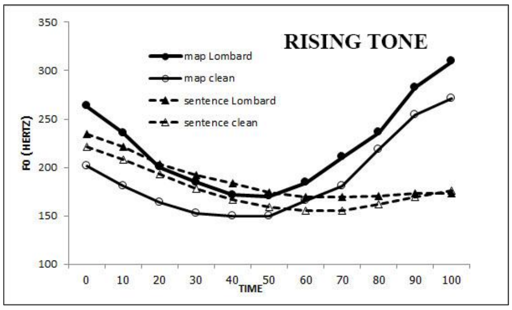


Figure 9: *F0 curves of rising tone across 10 normalized time points in clean vs. babble noise (Lombard) and with conversational partner (map) vs. with no partner (sentence).*

4. Discussion and Future Work

In general, our findings are in line with previous studies on Lombard speech [1-7]. Specifically, in babble noise condition, regardless of whether the tones were produced with or without conversation partner, F0 values for all tones were higher than in clean condition. Interestingly, Lombard Effect on Thai lexical tones was significantly increased (higher F0 for mid, falling, high, and rising tones) when conversation partner was present. On this latter point, it was consistent with what has been reported [9-11]. Thus, apart from noise condition, having a conversation partner can be another important factor for speakers to significantly adjust their speech. This is possibly because more of speaker's effort would require to obtain reactions from conversation partners when it comes to real and active communication situations.

From the current set of speech data, detailed analyses are being conducted on lexical tones in white noise condition and Lombard Effect on segmental durations with emphasis on adjustment of vowel duration among contrastively short and long vowels. Analysis of other acoustic correlates such as intensity and spectral tilt could also be valuable.

Another important contribution of this study is the design of map task (in Thai) and the procedure involved in the task. We believe that our map task could be beneficial for relevant studies in which speech eliciting in a natural conversation setting is required.

5. References

[1] Lau, P. "The Lombard effect as a communicative phenomenon," UC Berkley Phonology Lab Annual Report, 1-9, 2008.

[2] Anglade, Y. and Junqua, J-C., "Acoustic-phonetic study of-Lombard speech in the case of isolated-words," STL Research Reports, 129-135, 1990.

[3] Hazan, V. and Baker, R., "Acoustics-Phonetic characteristics of speech produced with communicative intent to counter adverse listening conditions," *J. Acoust. Soc. Am*., 2139-2152, 2011.

[4] Castellanos, A., Benedi, M. and Casacuberta, F., "An analysis of general acoustic-phonetic Features for Spanish speech produced with the Lombard effect," Speech Commun, 20:23–35, 1996.

[5] Zhao, Y. and Jurafsky, D., "The effect of lexical frequency and Lombard reflex on tone hyperarticulation," Journal of Phonetics, 37:231-247, 2009.

[6] Summers, W., Pisoni, D., Bernacki, Pedlow, R., and Stokes, M.,"Effects of noise on speech production: acoustic and perceptual analyses," *J. Acoust. Soc. Am.,* 84: 917-928, 1988.

[7] Patel, R. and Schell, K. W., "The influence of linguistic content on the Lombard effect," *J.Speech Lang. Hear.*, 51:209-220, 2008.

[8] Junqua, J-C., Fincke, S., and Field, K., "Influence of the Speaking Style and the Noise Spectral Tilt on the Lombard Reflex and automatic speech recognition," *In International Conference on Spoken Language Processin*g, 467-470, 1998.

[9] Lane, H., and Tranel, B., "The Lombard sign and the role of hearing in speech," W J. Speech Hear., 14:677-709, 1971.

[10] Fitzpatrick, M., Kim, K. and Devis, C., "The effect of seeing the interlocutor on auditory and visual speech production in noise," *In Proc. of International Conference on Auditory-Visual Speech Processing*, 31-35, 2011.

[11] Zeng, F.G. and Liu, S., "Speech Perception in individuals with auditory neuropathy," Journal of Speech, Language, and Hearing Research, 367-380, 2006.

[12] Kasisopa, B., Attina, V. and Burnham, D., "The Lombard effect with Thai lexical tones: an acoustic analysis of articulatory modifications in noise," Proceedings of Interspeech 2014, 15th Annual Conference of the International Speech Communication Association, Singapore, September 14-18, 2014.

[13] Onsuwan, C., Tantibundit, C., Saimai, T., Saimai, N., Chootrakool, P., and Thatphitthakkul, S., "Analysis of Thai tonal identification in noise," Proceeding of the 14th Australasian International Conference on Speech Science and Technology (SST). Sydney, Australia: Macquarie University, 2012.

[14] Mixdorff, H., Charnvivit, P., and Burnham, D. K., "Auditory–visual perception of syllabic tones in Thai," In E. Vatikiotis-Bateson, D. Burnham, & S. Fels (Eds.), Proceedings of the Auditory–Visual Speech Processing International Conference. Adelaide, Canada: Causal Productions. 3-8, 2005.

[15] Viethen, D. and Cox, F., "Designing a new map task (manuscript)," Sydney: Western Sydney University, 2010.

[16] Boersma, P. and Weenink, D., "Praat: Doing phonetics by computer [Computer Program] Version 5.3.57," retrieved 1 November 2013 from http://www.praat.org/

[17] Bapineedu, G., "Analysis of Lombard Effect Speech and its application in speaker verification for imposter detection," in Proc. of Interspeech 2013.

Whispered and Lombard speech: different ways to exaggerate articulation

Chris Davis and Jeesun Kim
The MARCS Institute, Western Sydney University
chris.davis/j.kim@westernsydney.edu.au

Abstract

When speaking in noise (Lombard speech) talkers exaggerate their articulation compared to speaking in quiet. The current study compared motion exaggeration in this speech style with whispered speech by measuring talker's lip and jaw, and eyebrow and head motion. Four talkers uttered sentences in quiet, noise or in whisper while their face and head movements were recorded with optical tracking. The results showed that both Lombard and whispered speech had movements of greater duration and amplitude than speech in quiet. For half the participants, whispered speech had greater motion than Lombard, whereas the other half showed the opposite pattern.

Index Terms: speech production, Lombard speech, whispered speech, exaggeration

1. Introduction

Speech style can have a considerable impact on the way that speech is articulated (e.g., hyper/hypo-articulation). Changes and variability in speech articulation are of interest to researchers from a range of disciplines (e.g., those researching speech production and recognition, speaker states, speech biometrics; multimodal speech, etc).

Exaggerated articulation is particularly interesting as it has been proposed to make speech more intelligible [e.g., 1]. Speech exaggeration has been observed in a number of different speech styles (e.g., clear speech, machine speech, foreigner directed speech, Lombard speech). In this study, we contrast a well-studied speech style that is known to produce hyper-articulation (speaking in noise, Lombard speech) with a less studied style that has also been suggested to exhibit hyper-articulation (whispering).

Comparing whispered and Lombard speech is interesting because part of the exaggerated articulation of Lombard speech may be simply due to speaking loudly, whereas for whispered speech it may be due to a strategy to produce clear visual speech. If this were the case, then whispered and Lombard speech may show slightly different motion patterns. For instance, exaggerated jaw and lip motion may be a necessary concomitant of Lombard speech but this may not be the case for eyebrow and head motion that are less directly coupled to speech articulation (although see [9]).

Before describing the current study in more detail, it is useful to note some aspects of whispered speech that were considered in planning the current study. Whispered speech is typically produced with an open glottis and without voicing. Whispered speech has been classified into low and high energy whisper [2]. Low-energy whisper is typically produced in situations where interlocutors wish to maintain local intelligibility but deliberately attempt to reduce the perceptually salience of speech to others (for example, whispering in a library or in a formal or solemn setting). High-energy whisper (also known as 'stage whisper') is produced in order to be intelligible at a distance and is characterized by greater pulmonic force and air flow. In this type of whisper, the voice is still not produced with active vocal cord vibration but occasional passive vibration can occur.

Studies of whispered speech have mainly concentrated upon characterizing the vocal mechanisms of whisper production (e.g., glottal configuration, [3; 4]) or examining the perception of whisper [5]). The few studies that have examined articulation have typically looked at the production of segments (e.g., /p/, /b/ or single vowels). For example, using an intra-oral pressure measure, Schwartz [6] found significantly longer bilabial closure and significantly greater whispered durations for /p/ and /b/ but not /m/. Using a spectrographic analysis method Parnell et al [7] found that closure and constriction durations for /t/, /s/, and /z/ were significantly longer in whispered vowel environments compared to voiced ones. Higashikawa et al [8] used a video-based analysis of reflective markers positioned on both the upper and lower lip and to the protuberance of the mandible. These authors found differences between lip movements for bilabial plosives during normal and whispered speech. That is, lip opening was significantly faster when whispering /b/ compared with whispering /p/ or non-whispered /b/. In summary, it has been demonstrated that at least for some segments, whispered production speech may have greater duration and involve greater articulatory motion.

The current study examined high energy whisper (i.e., the interlocutor was separated from the speaker by several meters due to restriction imposed by the recording method). In this regard we note that Solomon et al [2] showed that low-energy and high-energy whispering could be differentiated by supraglottal constriction and to lesser extent by vocal-fold adjustments, (although individual differences tended to be considerably larger than any systematic effects due to the type of whisper). It also seems plausible, given their different functional roles, that the degree of visual articulation will differ between the two types of whisper.

To examine the articulator effects of whisper and Lombard speech, we used active infrared optical tracking (Optotrak) and measured articulatory motion from the lips and jaw (movements directly related to speech articulation as in Higashikawa et al [8]), as well as eyebrow and rigid head motion. Also, we examined motion at the sentence (unlike [8]) rather than segment level. This was done to obtain a more global impression of the effects of exaggeration on articulation. We also examined speech production in quiet in order to determine a baseline against which hyper-articulation could be judged.

2. Method

2.1. Participants

Four people participated in the experiment (3 males, 1 female). All were native speakers of English (one British, two Australian and one American); ages ranged from 32 to 54 years.

2.2. Materials

10 sentences selected from the 1965 revised list of phonetically balanced sentences (Harvard Sentences, [10]).

2.3. Apparatus

Two Northern Digital Optotrak machines were used to record the movement data. The configuration of the markers on the face and head rig is shown in Figure 1.

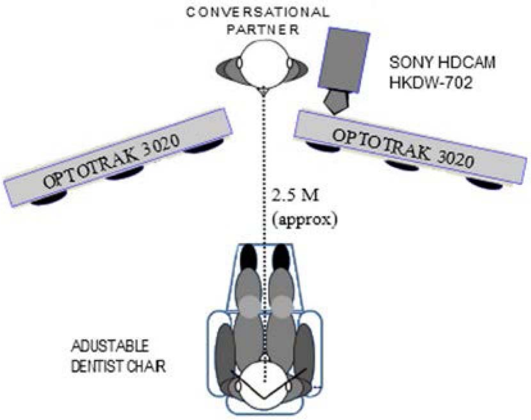


Figure 1: *The layout of the capture session. The participant whose head and face movement were recorded sat in an adjustable chair facing a "conversational partner" who stood behind the two NDI Optotrak 3020 machines.*

2.4. Procedure

Each session began with the placement of the movement sensors (see Figure 1) during which time participants were asked to memorize the ten sentences to be spoken. Each participant was recorded individually in a session that lasted approximately 90 minutes. Participants were seated in an adjustable dentist chair in a quiet room and were asked to say aloud ten sentences (one at a time) to a person who was directly facing them at a distance of approximately 2.5 meters (See Figure 1). The participant then repeated the ten sentences.

This basic procedure was repeated several times, once for each speech mode (in quiet, whispered, Lombard). Lombard speech was induced by participants speaking while hearing multi-talker babble through a set of ear phones (at approximately 80 dB SPL, a similar level to [1]). For whispering, participants were instructed to whisper the sentences at a level judged loud enough for the conversational partner to hear.

2.5. Data processing

Non-rigid facial and rigid head movement was extracted from the raw marker positions. The data were recorded at a sampling rate of 60Hz. To guard against the over-representation of particular marker configurations, a movement threshold quantification procedure was employed to keep only the frames that were sufficiently different. *Guided* Principle Component Analysis (gPCA, see [11]) was used to reduce the dimensionality of the data; this procedure uses linear decomposition to constrain the PCA to motion planes that are relevant to articulation using *a priori* defined markers (jaw, lips, eyebrows).

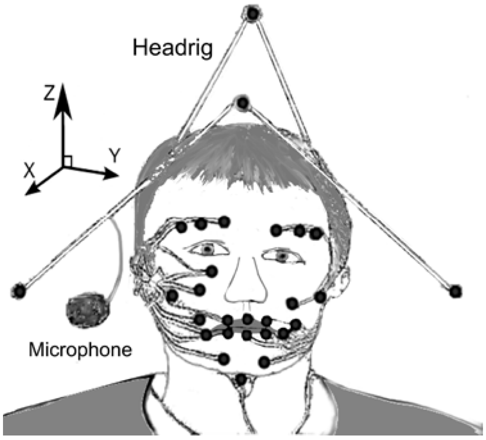


Figure 2: *The location of the 24 optical sensors on the face (the size of the sensors have been exaggerated for clarity). Four additional sensors were positioned on a head-rig to measure rigid movements around the centre of rotation.*

For the current study we report the following gPCA components. Direct speech articulation: Jaw: Jaw Opening; Jaw Protrusion; Mouth: Lip Opening; Lip Rounding. Indirect articulation: Eyebrow: Eyebrow Raising; Eyebrow Pinching. Head translation: Forward / Backwards; Up / Down.

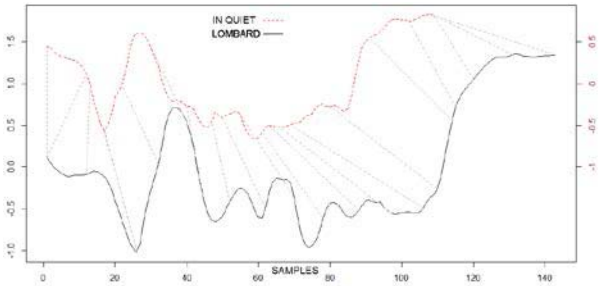


Figure 3: *An example of two time series showing the contribution of a guided principle Component (here Lip opening) for speech 'in quiet' and 'in noise'. Also shown is the DTW between them (the vertical axis represents cm, left scale in noise condition).*

In order to quantify the degree of hyper-articulation of the production component we used Dynamic Time Warping (DTW) [12]. Dynamic time warping (DTW) is a procedure that provides a measure of comparison between series of data points (inherent distance or warping cost). For example, DTW can expand or compress one time series to resemble another one and by summing the distances of individually aligned elements an inherent distance between the two can be computed (Figure 3).

We compared the warping cost of time-series of the contribution of the gPCs over each utterance for the speech in quiet condition compared to either the Lombard or the whispered speech conditions. Note that these time-series were mean-centered to avoid the effect of off-sets.

In order to have an index of the power of the contribution of a PC, we used the standard deviation of the mean-centered time-series.

3. Results

Figure 4 shows the mean inherent distance scores (warping costs) for whispered or Lombard speech compared with speech produced in quiet for the following motion types: jaw (opening and protrusion) and lips (opening and rounding), as well as eyebrow (up-down and pinched) and rigid head motion (forward–back and up-down).

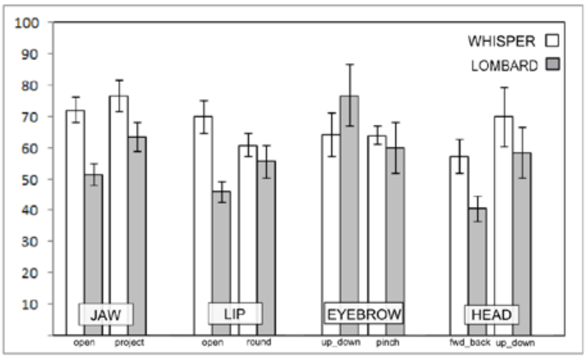


Figure 4: *Mean inherent distance scores (warping cost in arbitrary units) for the different motion types as a function of speech type. Error bars show Standard error.*

It is clear from the magnitudes (and error bars) of the warping cost scores that both the whisper and Lombard speech motion curves differed from those for speech in quiet (the baseline against which the distance scores were determined).

The warping cost scores were analyzed using a linear mixed model (LMM) analysis (random-intercept) using the LmerTest package to approximate degrees of freedom [13]. This analysis indicated that the effect of speech type (whisper vs. Lombard) was significant, F(1,626) = 9.877, $p < 0.01$. This effect was not significant if random slopes were included in the model, indicating a possible interaction effect with participants (although [14] have also argued that maximal models can be unduly conservative).

Figure 5 shows the average warping cost scores for all motion types as a function of participants. As can be seen in the figure, there was considerable variation across participants in terms of whether there was more motion in articulating whispered speech or Lombard speech. Indeed, the interaction between the speech type and participant was significant, t = 4.564, $p < 0.05$.

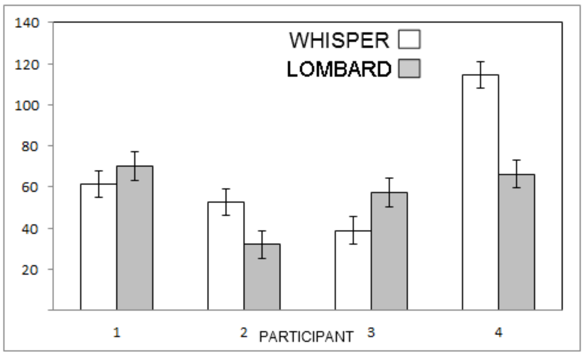


Figure 5: *Mean inherent distance for all motion types as a function of speech type and participant. Error bars show Standard error.*

Figure 6 shows a comparison of distance scores (relative to quiet speech) for whisper and Lombard speech broken down by eyebrow and head motion and jaw and lip motion. Across each individual participant, the pattern of motion for the jaw and lips (motion closely related to speech articulation) was similar to the pattern for the eyebrows and head. This was the case even though there was variability across participants as to whether Lombard speech had more motion than whispered speech or vice-versa.

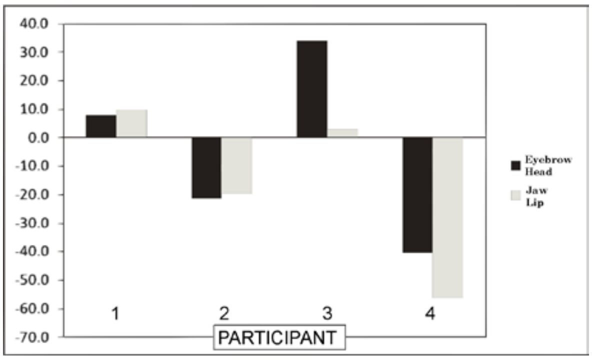


Figure 6: *Mean inherent distance scores for Lombard minus whispered speech (negative means larger warping costs for whispered speech) for Eyebrow and Head motion (black) and Jaw and lip motion (grey) for the participants (1 – 4)*

To determine whether there was a difference in the amount of the contribution of the gPCs, the standard deviation of the mean-centred time-series was calculated as a measure of the PC contribution. This value represents the amount to which the PC score deviated around the mean (which represents the initial configuration). The data for each of the motion PCs are shown in Figure 7.

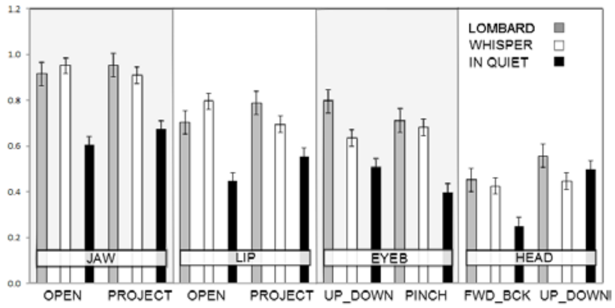


Figure 7: *Mean of the standard deviations (in arbitrary units) of the different motion PCs as a function of speech type. Error bars show Standard error.*

As can be seen in the figure, there was a more power (greater deviation) for the Lombard and whispered speech motion PCs compared to speech uttered in quiet. Two LMMs (random slopes, intercepts) were conducted to determine whether there was greater variation (amplitude) for the whispered and Lombard speech conditions each compared to the speech in quiet one. The LMM for Lombard versus in quiet was significant, F(1,4) = 10.343, $p < 0.05$; as was the LMM for whispered versus in quiet, F(1,3) = 63.8, $p < 0.01$. There was no difference between the Lombard and whispered speech scores, F(1,3) = 0.3125, $p > 0.05$.

4. Discussion

Speech related mouth and jaw articulation, along with eyebrow and head motion was measured for three speech styles (speech in quiet, in noise and whisper). Two methods for quantifying differences in motion of guided principle components were used. The inherent distance (warping cost) between motion PC curves (as given by DTW) for speech in quiet and two the other speech styles indexed temporal exaggeration. The other measure, the deviation of PCs from a centred mean value, indexed the power (amplitude) of the motion contribution. The results of both analyses support the claim that high-energy whispered speech is hyper-articulated compared to speech produced in quiet.

The results concerning whether the degree of whispered speech motion differed from that produced when speaking in noise (Lombard speech) were less clear. The results taken over all speakers suggest that whispered speech may exhibit more hyper-articulation than Lombard speech. This finding is consistent with the proposal that speakers can strategically employ visual speech to aid communication when there is a barrier to auditory communication [15]. However, there was considerable individual variation in this pattern, with greater motion for whispered speech only found for two of the four participants (and two showing the opposite pattern). There was also no clear difference in the patterning of motion of the lips and jaw (motion tied to speech articulation) compared with that of the eyebrow and head. Further investigation is warranted.

Finally, it is interesting to point out practical implications of the current findings. For example, the finding that there is hyper-articulation for (high-energy) whispered speech (that, for some people, is even greater than that shown for Lombard speech) is relevant to a range of speech research topics: from the examination of methods of reliably detecting speech from face motion [16] to auditory-visual speech biometric systems, where whisper may be a useful speech style to use as hyperarticulation may be more distinctive.

5. Acknowledgements

We acknowledge the support of the Australian Research Council grant DP130104447.

6. References

[1] Junqua J-C. (1996). The influence of acoustics on speech production: A noise-induced stress phenomenon known as the Lombard reflex. Speech Communication, 20, 13-22.

[2] Solomon, N. P., McCall, G. N., Trosset, M. W., & Gray, W. C. (1989). Laryngeal configuration and constriction during two types of whispering. Journal of Speech, Language, and Hearing Research, 32(1), 161-174..

[3] Esling, J. H. & Harris, J. G. 2003. An expanded taxonomy of states of the glottis. Proceedings of the 15th International Congress of Phonetic Sciences, 1049-1052. Barcelona, Spain:UAB.

[4] Mills, T. I. P. (2009). Speech motor control variables in the production of voicing contrasts and emphatic accent. Doctoral dissertation, University of Edinburgh.

[5] Tartter, V. C. (1994). Hearing smiles and frowns in normal and whisper registers. Journal of the Acoustical Society of America, 96(4), 2101-2107.

[6] Solomon, N. P., McCall, G. N., Trosset, M. W., & Gray, W. C. (1989). Laryngeal configuration and constriction during two types of whispering. Journal of Speech, Language, and Hearing Research, 32, 161-174.

[7] Parnell, M., Amerman, J. D., & Wells, G. B. (1977). Closure and constriction duration for alveolar consonants during voiced and whispered speech. Journal of the Acoustical Society of America, 86, 1678-1683.

[8] Higashikawa, M., Green, J. R., Moore, C. A., & Minifie, F. D. (2003). Lip kinematics for/p/and/b/production during whispered and voiced speech. Folia phoniatrica et logopaedica: official organ of the International Association of Logopedics and Phoniatrics (IALP), 55(1), 17.

[9] Davis, C., Kim, J., Grauwinkel, K., & Mixdorff, H. (2006, May). Lombard speech: Auditory (A), Visual (V) and AV effects. In Proceedings of the Third International Conference on Speech Prosody (pp. 248-252).

[10] Harvard sentences: Appendix of: IEEE Subcommittee on Subjective Measurements IEEE Recommended Practices for Speech Quality Measurements. IEEE Transactions on Audio and Electroacoustics. 17, 227-46, 1969.

[11] Maeda, S. (2005). Face models based on a guided PCA of motion capture data: Speaker dependant variability in /s/ - /z/ contrast production. ZAS Papers in Linguistics, 40, 95-108.

[12] Giorgino T (2009).Computing and Visualizing Dynamic TimeWarping Alignments in R: The dtw Package." Journal of Statistical Software, 31(7), 1{24. URL http://www.jstatsoft.org/v31/i07/

[13] Kuznetsova, P. B. Brockhoff, R. H. B. Christensen (2013). lmerTest: Tests for random and fixed effects for linear mixed effect models (lmer objects of lme4 package). R-Version:1.1-0. http://cran.rproject.org/web/packages/lmerTest/index.html.

[14] Matuschek, H., Kliegl, R., Vasishth, S., Baayen, H., & Bates, D. (2015). Balancing Type I Error and Power in Linear Mixed Models. arXiv preprint arXiv:1511.01864.

[15] Fitzpatrick, M., Kim, J., & Davis, C. (2011). The effect of seeing the interlocutor on speech production in different noise types. In Twelfth Annual Conference of the International Speech Communication Association.

[16] Denby, B., Schultz, T., Honda, K., Hueber, T., Gilbert, J. M., & Brumberg, J. S. (2010). Silent speech interfaces. Speech Communication, 52(4), 270-287.

Revisiting the interlanguage speech intelligibility benefit

Chang Shu, Ian Wilson, Jeremy Perkins

University of Aizu, Japan

m5192107@u-aizu.ac.jp, wilson@u-aizu.ac.jp, jperkins@u-aizu.ac.jp

Abstract

Bent and Bradlow (2003) first discovered evidence for an interlanguage speech intelligibility benefit, essentially non-native listeners finding similar-L1 non-native speech equally or more intelligible than native speech. We have refined their method by using 14 speakers from 7 languages (English, Chinese, Hindi, Japanese, Korean, Russian, and Vietnamese) and using reaction time (RT) to accented speech as a more sensitive measure of intelligibility than transcription tasks. Non-native participants (15 Japanese, 9 Chinese, and 6 Vietnamese) had significantly faster RTs to same-accent speakers than to other L2 speakers. L1 English participants had faster RTs to L1 English speakers than to L2 speakers.

Index Terms: reaction time, interlanguage speech intelligibility benefit, English, Japanese, Chinese, Vietnamese, L2 speech

1. Introduction

Bent and Bradlow [1] did an intelligibility study using noise-embedded spoken data from 1 native English speaker, 2 Chinese speakers of English, and 2 Korean speakers of English. The listeners who transcribed the spoken data were from a variety of first-language backgrounds: monolingual English (n=21), Chinese (n=21), Korean (n=10), and other (n=12) – where “other” meant one of Bulgarian, Dutch, French/Douala, German, Greek, Hindi, Japanese, Serbian, Spanish, or Tamil. The following two findings were especially interesting: (i) non-native listeners found high-proficiency non-native speakers of the same L1 equally as intelligible as they did native speakers, something they called the “matched interlanguage speech intelligibility benefit”, and (ii) even when listeners were not of the same L1 as the speakers, they found high-proficiency non-native speakers equally as intelligible as they did native speakers, something they called the “mismatched interlanguage speech intelligibility benefit.”

Stibbard and Lee’s [2] replication of [1] did not show results that supported [1]; one of their findings, for example, was that native speakers were not more intelligible than non-native speakers even to their fellow native listeners. Some procedures in [2] were different from that of [1], though, making it more difficult to directly compare the results. For example, presentation of stimuli was randomised in [2], eliminating a possible familiarity effect in [1], and the sentences in [2] were not embedded in noise (although no ceiling effect occurred).

Xie and Fowler [3] investigated the intelligibility of native and Mandarin-accented English speech for native English and native Mandarin listeners, specifically at the acoustics of stop voicing. They also made a distinction between interlanguage speech intelligibility benefit for listeners and that for speakers.

One drawback, however, of both [1] and [2] is that they each used only five speakers: one native speaker of English and two non-native speakers from each of the respective languages studied (Chinese and Korean in [1] and Saudi and Korean in [2]). And in [3], only one native English speaker and one Mandarin speaker were used. Thus, as pointed out in [2], results could have been greatly affected by idiosyncrasies of the speakers’ speech rather than a given foreign accent in general. The same limitation is true of a study by Chen [4], who investigated 29 native and non-native listeners’ perceptual judgements of the intelligibility of Chinese-accented speech. She had one Cantonese speaker from HK and one Mandarin speaker from Taiwan, although the listeners were from various language backgrounds. On a dictation task measuring intelligibility, all groups scored higher for the Mandarin accent than for the Cantonese accent, except for Cantonese listeners. This suggests a benefit for listeners listening to speech produced with their own accent.

All four of the studies discussed above used a dictation task to measure intelligibility. Using a more sensitive measurement such as reaction time (RT) can improve the sensitivity of an intelligibility test [5]. Measurements of RT have been used in speech intelligibility tests for over 50 years [6], and speech spoken with a foreign accent is indeed less intelligible (i. e., people take longer to react to it – meaning RTs are longer) than native speaker speech, at least to a native listener [7].

Studies have found that listeners could transcribe utterances perfectly, even though they rated the speakers as heavily accented, indicating that accent does not necessarily result in reduced intelligibility [8], even though it could still increase RT. Although RTs to foreign-accented speech are initially slower than to native speech, it has been shown that listeners can very quickly adapt (in as few as 2–4 utterances) [7]. In that study, native (Tucson, AZ) English speakers’ RTs were measured as they listened to the speech of (i) a native speaker of English, (ii) a non-native Spanish-accented speaker of English, and (iii) a non-native Chinese-accented speaker of English. Results showed that RT to non-native speech was slower than RT to native speech, but that the difference diminishes within 1 minute of exposure.

RT is also slower when listening to a dialect of one’s native language (L1) that is different from one’s own dialect [9]. Native French listeners’ RTs to various dialects of French were measured, and they found a significant cost to listening to a different dialect of one’s L1 – a 30 ms delay in word identification. Unlike [7], they used multiple speakers for each of the dialects of the L1. In a follow-up study [10], it was found that accent changes cause a temporary perturbation in RTs and that this delay in word identification does not disappear after repeated exposure to the same accent.

In this research, we tested for the interlanguage speech intelligibility benefit reported in past research [1], but we used a total of 14 speakers of 7 languages, 30 listeners, and used arguably a more sensitive measurement of intelligibility: the RT of participants identifying one of two images on the screen – the image corresponding to the audio prompt.

2. Method

2.1. Participants

Thirty-five participants took part in the RT experiment (26 male, 9 female), including 5 native English speakers, 15 native Japanese speakers, 9 native Mandarin Chinese speakers and 6 native Vietnamese speakers. In the English L1 group, there were 3 Canadians, 1 American, and 1 New Zealander. In the Japanese L1 group, there were 2 graduate students and 13 undergraduate students. In the Chinese and Vietnamese groups, all participants were either graduate students or working full time. All L2 participants had studied English for at least 7 years, although to different proficiency levels. Most of them (27 out of 30) had taken the TOEIC English proficiency test in the last 2 years, with scores ranging widely from 265 to 960. All participants were right-handed, except one Mandarin Chinese speaker.

2.2. Stimuli

Eight pairs of words were selected as stimuli, and each word in a pair shared similarities. Table 1 shows the list of stimuli, all of them nouns. The first number is the frequency ranking in the Corpus of Contemporary American English (CCAE) [11], where 1 = the most frequent word in American English, and the column with a two-digit code shows the approximate grade (Gr) that the word is learned in public Japanese schools (J = Junior High; S = Senior High). Note that both words in a given pair were approximately balanced in frequency, and all were common English words.

Table 1: *Stimuli list with frequency and grade level*

Pair	Stimulus 1			Stimulus 2		
No.	Word	CCAE	Gr	Word	CCAE	Gr
1	food	367	J2	foot	381	J2
2	glass	823	J1	gas	1026	J3
3	nose	1748	J2	snow	1795	J2
4	shape	1273	S1	shoe	1430	J2
5	cat	1788	J1	hat	2033	J2
6	flight	1302	J3	fight	1573	J3
7	lake	2204	J2	cake	2563	J1
8	wall	572	J1	ball	915	J2

Combinations of simultaneous visual and audio prompts were used to present the stimuli. Eight image pairs were created, with the left words in Table 1 on the left and the right words on the right (e. g. in Image 1, a picture of food on the left and foot on the right). All 16 stimuli were inserted into the carrier sentence “The picture you should choose is ______.” and they were read and recorded by 14 speakers in a light-type soundproof booth [12]. The 14 speakers were all University of Aizu professors; 2 speakers from each of 7 different countries (Canada, China, India, Japan, Korea, Russia, Vietnam).

A total of 224 audio-visual stimuli (8 image pairs × 2 words per pair × 14 speakers) were created, and then divided into 2 blocks. In Block 1, 1 randomly-chosen word from each pair was read by 1 randomly-chosen speaker from each of the 7 countries. In Block 2, the other word from the same pair was read by the other speaker from each of the 7 countries.

So, each block contained 112 stimuli and each participant (listener) was asked to listen to one of the blocks. For every participant from a given L1, we alternated the choice of block, so the first Japanese participant did block 1, the second block 2, the third block 1, etc.

2.3. Data collection

Before doing the RT experiment, participants filled out a questionnaire asking about handedness, English learning experience, standardized test scores, etc. All participants were offered money to participate, but a few of them refused to be paid. The experiment was conducted in the same soundproof booth as the stimuli were recorded in, ensuring a quiet environment.

Before starting the actual experiment, each participant had brief training. In the training session, participants were requested to respond as quickly as possible to 3 pairs of stimuli, which had the same kind of combination of picture and sound (but were different words from the ones used in the actual experiment). The training stimuli were all spoken by a Colombian speaker of Spanish, and his voice was not used outside of training. Participants adjusted the volume of the sound in their headphones to a comfortable level.

The input device used was an Xbox 360° controller joystick, and participants had to press the left or right button with their left or right index finger, as soon as they determined which picture matched the word they heard. The order of stimuli presentation was randomised by E-Prime 2 software running on an HP EliteBook 8570w laptop computer. As RT was measured from the beginning of the target word (the last word of the sentence) and the longest sound file was about 4.4 seconds, we allowed the images to be displayed after the audio prompt had stopped until a response was given, or until a response deadline of 8 seconds was reached.

2.4. Data analysis

We first analyzed the RT data generated by E-Prime 2 and found that participants had answered incorrectly in 7.2% of trials. All such trials were eliminated from further analysis. No single participant had an overall error rate greater than 20%. Responses from five individual stimuli were excluded from analysis due to error rates greater than 40% across all participants. Four of these stimuli were of non-native English speakers saying “food” (two Chinese speakers: 84% and 69% error rates; one Japanese speaker: 50%; and one Vietnamese speaker: 58%). The remaining stimulus was of a non-native speaker of Russian saying “foot” (63%). “Food” and “foot” have vowel length and quality differences that are difficult for non-native speakers, both in production and perception, possibly causing high error rates.

Also, responses were excluded that were more than 3 times the median average deviation (MAD) from the median RT, calculated separately for each listener. MAD was used rather than SD since it is less influenced by outliers, following [13]. Following [14], we calculated MAD per listener, rather than over the entire data set, since RTs had a large inter-speaker variation. A total of 502 out of 3920 responses were excluded (12.8%), leaving 3418 responses for analysis. Of these 502 responses, 313 (8.0%) were excluded because of high error rates, and 189 (4.8%) were excluded because they exceeded the threshold of 3 times the MAD from the participant median.

The lme4 package [15] and the lmerTest package [16] were used in R [17] to perform a linear mixed effects analysis of the relationship between RT and the factors Participant Language (PL), Speaker Language (SL), and Language Relation (LR). Incorrect responses were omitted from the analysis. The factor LR was coded according to whether the speaker and listener spoke the same native language or not (LR has two levels: ”same” or ”different”). The final model had three fixed effects (PL, SL, and LR) with no interaction terms.

As random effects, intercepts were included for Word and

Table of Contents for this manuscript

Participant, but we did not include random slopes. The linear mixed model fit was performed via the REML method, with rival models assessed by likelihood ratio tests, incrementally removing fixed effects. Alternate models with interaction terms added were assessed as well, using the maximum-likelihood method. To assess whether a fixed effect parameter estimate was significant, t-tests were performed using Satterthwaite approximations to degrees of freedom. Type-III F-tests were used to assess the significance of a given fixed effect.

Since speaking rate differed across speakers, and since we did not want to artificially alter any stimuli, we checked Speaker as a random effect and found that it was insignificant in an analysis of random effects (p = 1). As a result, Speaker was excluded as a random effect in the final model.

3. Results

Mean RTs categorized by native English status of speaker and listener and by language relation (LR) are summarized in Table 2. The fastest RTs were by native English participants listening to native English speakers (651 ms), and the slowest RTs were by non-native participants listening to other non-native participants who did not share the same L1 (851 ms). Linear mixed effects analysis details are given in the following paragraphs.

Table 2: *Mean RTs (ms) categorized by native English status of participant & speaker, and by language relation*

Part. L1	Spkr L1	RT (ms)	S.E.	n
Engl.	Engl.	651	28	77
	Non-Engl.	681	8.4	2056
Non-Engl.	Engl.	807	9.5	430
	Non-Engl. (same)	811	17	430
	Non-Engl. (diff.)	851	16	425

Significant effects on RT for speaker language, $F(6, 3381) = 5.6$, $p < 0.01$, participant language, $F(3, 31) = 8.3$, $p < 0.01$, and LR, $F(1, 3381) = 5.3$, p = 0.02, were discovered. The model intercept estimate, using the condition with native English listeners and native English speakers (LR = ”same”) yielded an RT of 615 ms (SE = 86, $p < 0.01$). A significant positive effect was found on RT when listener and speaker had different native languages ($\beta = 31.748$, SE = 13.8, p = 0.02).

In addition, all participants responded more quickly to stimuli spoken by a native English speaker than to non-native speakers of Chinese ($\beta = 74.8$, SE = 16.7, $p < 0.01$), Hindi, ($\beta = 49.1$, SE = 16.6, $p < 0.01$), Japanese ($\beta = 43.4$, SE = 16.9, p = 0.01), Korean ($\beta = 70.6$, SE = 16.6, $p < 0.01$), and Russian ($\beta = 46.3$, SE = 16.4, $p < 0.01$); however the RTs were not significantly slower in responses for the Vietnamese speakers ($\beta = 13.5$, SE = 16.3, p = 0.41).

Mean RTs are listed by native language of speaker (columns) and participant (rows) in Table 3. In general, native English participants responded faster to all stimuli regardless of the speaker’s native language (relative to Chinese participants: $\beta = 286$, SE = 102, $p < 0.01$; and Vietnamese participants: $\beta = 362$, SE = 110, $p < 0.01$); however, there was no significant difference between the RTs of native English and native Japanese participants ($\beta = 13.9$, SE = 94.1, p = 0.88).

There was no evidence of significant interactions between LR and SL, or between LR and PL (model AIC with interaction = 47854; model AIC without interaction = 47850; $\chi^2(3) = 2.56$, p = 0.46 for both likelihood ratio tests). This indicates that RTs were faster if the listener and speaker spoke the same native language, independent of the native language.

Table 3: *Mean RT (ms) by speaker L1 (columns) and participant L1 (rows); letters are the languages English, Chinese, Hindi, Japanese, Korean, Russian, and Vietnamese*

Part.	Speaker L1							
L1	E	C	H	J	K	R	V	All
E	651	704	659	680	701	690	656	677
C	933	974	957	1013	997	974	910	965
J	666	707	723	657	710	690	688	691
V	972	1077	1066	1001	1092	1048	969	1032

Of the 16 stimuli used, all had mean accuracy rates higher than 90%, except for “food” and “foot”, which had rates of 71.8% and 75.1% respectively. Among correct responses, these two words also had significantly longer RTs than other words, and were the only words to have mean RTs higher than 1000 ms across all speakers. Notably, “shape” and “shoe” had relatively high RTs (958 ms and 968 ms respectively), perhaps because they were the only other pair with identical syllable onsets.

English proficiency (TOEIC test scores) were collected from non-native English speaking participants. Proficiency did not significantly lower RT; in fact, a small but significant effect was seen where proficiency correlated positively with RT ($\rho = 0.058$, t = 3.034, $p < 0.01$). For the purposes of the linear mixed model analysis, it was decided to exclude the effect of proficiency, and instead include participant as a random intercept. The correlation between RT and proficiency is illustrated in Figure 1.

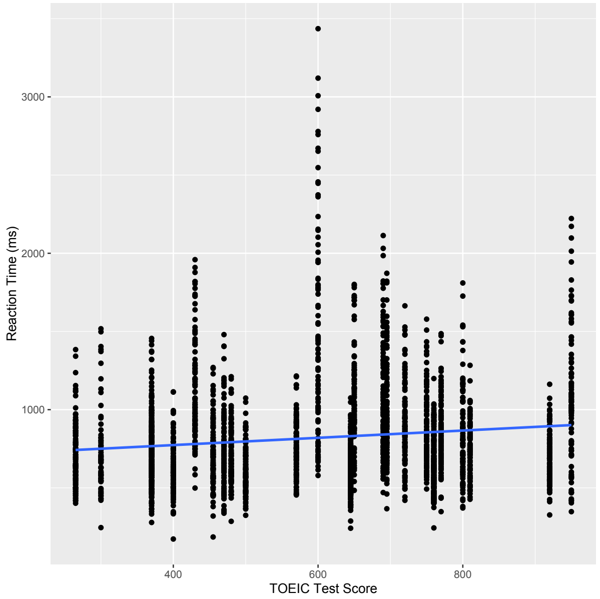


Figure 1: *RT (ms) plotted against English proficiency (TOEIC) test scores of participants*

4. Discussion

Using a more sensitive measure of intelligibility than transcriptions of speech in noise, we found that non-native listeners find speech significantly more intelligible if it is produced by speakers from the same accent group rather than a different accent group (e. g., Japanese speakers find Japanese-accented English more intelligible than other-language-accented English).

Non-native listeners even find their own accent group's L2 English more intelligible than native English, something that agrees with results found in [1]. Note that in [1], this was only found when the speaker was high proficiency, though, but the speakers we used had a great variety of degree of accent. For example, the Japanese speakers in our study are both low proficiency speakers of English, and yet Japanese listeners had faster RTs to their accented English than to native English.

All 14 speakers were professors at the same university, and all L2 participants except one were students at that university. It is certainly possible that some students recognised some voices and this may have affected the RTs. On the other hand, most of the graduate students who have a research supervisor of the same L1 speak to that professor in the L1 instead of in English. One of the Vietnamese participants was surprised to learn (after the experiment) that his research supervisor with whom he meets weekly was one of the 14 speakers. It is not so surprising, though, given the fact that that student and his supervisor converse in Vietnamese, not English, during their meetings.

Instead of simply using the target word in our audio prompts, we decided to embed those words in a carrier sentence. The reason for this is because it gives participants about 2 to 3 seconds to adapt to the speaker's accent and it has been shown that adaptation can be done in a short time [7]. The carrier sentence may also help to minimise RT perturbations that were found in [10] when the speaker changes from trial to trial. When a student is listening to a lecture in accented English, they have ample time to adapt to the speaker's accent, so using a carrier sentence helped to make the situation more realistic. It should be pointed out, though, that the participants' RTs may have been affected by expectations resulting from the pronunciation of the carrier sentence. One participant, a native speaker of English, told us that he was more likely to press the joystick button sooner when the carrier sentence was spoken by another native speaker because he had more confidence that a pronunciation error would not be made.

It is somewhat surprising that Japanese participants had RTs that were not significantly different from native English participants. One possibility for this is age; the native English participants were in their 30s and 40s, except one in his mid-20s, while the Japanese participants were all undergraduates about 21 years old.

It should be unsurprising that no strong correlation was found between RTs and the English proficiency scores of the participants, because the words were specifically chosen to be very common words learned before the end of the first year of senior high school. The weak positive correlation may be due to the fact that the graduate student participants, who were older (thus subsequently slower?) than the undergraduates, had higher TOEIC scores.

5. Conclusions and future work

In conclusion, L2-English participants (Japanese, Chinese, and Vietnamese) had significantly faster RTs to same-L1 speakers' English than to different-L1 speakers' non-native English, a type of matched interlanguage speech intelligibility benefit [1]. Native English listeners had faster RTs when listening to native English speakers than when listening to L2 speakers of English.

In the future, we would like to see how effective it would be to train students to perceive accented speech (e. g. crosslinguistic phonetic and phonological differences). We can then measure effectiveness by comparing pre- and post-training RTs.

6. Acknowledgements

This research was supported by 2016 University of Aizu Competitive Research Funding awarded to Ian Wilson. The authors are grateful to all 15 University of Aizu professors whose voices were recorded for stimuli, and to all the participants.

7. References

[1] T. Bent and A. R. Bradlow, "The interlanguage speech intelligibility benefit," *Journal of the Acoustical Society of America*, vol. 114, no. 3, pp. 1600–1610, 2003.

[2] R. M. Stibbard and J.-I. Lee, "Evidence against the mismatched interlanguage speech intelligibility benefit hypothesis," *Journal of the Acoustical Society of America*, vol. 120, no. 1, pp. 433–442, 2006.

[3] X. Xie and C. A. Fowler, "Listening with a foreign-accent: The interlanguage speech intelligibility benefit in mandarin speakers of english," *Journal of Phonetics*, vol. 41, pp. 369–378, 2013.

[4] H. C. Chen, "Judgments of intelligibility and foreign accent by listeners of different language backgrounds," *The Journal of Asia TEFL*, vol. 8, no. 4, pp. 61–83, 2011.

[5] C. Gooskens, "Experimental methods for measuring intelligibility of closely related language varieties," in *The Oxford Handbook of Sociolinguistics*, R. Bayley, R. Cameron, and C. Lucas, Eds. Oxford, UK: Oxford University Press, 2013, pp. 195–213.

[6] M. H. L. Hecker, K. N. Stevens, and C. E. Williams, "Measurements of reaction time in intelligibility tests," *Journal of the Acoustical Society of America*, vol. 39, no. 6, pp. 1188–1189, 1966.

[7] C. M. Clarke and M. F. Garrett, "Rapid adaptation to foreign-accented English," *Journal of the Acoustical Society of America*, vol. 116, no. 6, pp. 3647–3658, 2004.

[8] M. J. Munro and T. M. Derwing, "Foreign accent, comprehensibility, and intelligibility in the speech of second language learners," *Language Learning*, vol. 45, no. 1, pp. 73–97, 1995.

[9] C. Floccia, J. Goslin, F. Girard, and G. Konopczynski, "Does a regional accent perturb speech processing?" *Journal of Experimental Psychology: Human Perception and Performance*, vol. 32, no. 5, pp. 1276–1293, 2006.

[10] C. Floccia, J. Butler, J. Goslin, and L. Ellis, "Regional and foreign accent processing in English: Can listeners adapt?" *Journal of Psycholinguistic Research*, vol. 38, pp. 379–412, 2009.

[11] M. Davies. (2008) The corpus of contemporary american english: 520 million words, 1990-present. [Online]. Available: http://corpus.byu.edu/coca/

[12] Kawai. (2015) Kawai musical instruments manufacturing co. catalog. [Online]. Available: http://www.kawai-os.co.jp

[13] C. Leys, C. Ley, O. Klein, P. Bernard, and L. Licata, "Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median," *Journal of Experimental Social Psychology*, vol. 49, no. 4, pp. 764–766, 2013.

[14] R. Ratcliff, "Methods for dealing with reaction time outliers," *Psychological Bulletin*, vol. 114, no. 3, pp. 510–532, 1993.

[15] Douglas M. Bates and Martin Maechler and Ben Bolker, *lme4: Linear mixed-effects models using S4 classes*, R package version 1.1-12, 2016.

[16] A. Kuznetsova and P. B. Brockhoff and R. H. B. Christensen, *lmerTest: Tests in Linear Effects Models*, R package version 2.0-32, 2016.

[17] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2015, ISBN 3-900051-07-0. [Online]. Available: http://www.R-project.org/

Table of Contents for this manuscript

Speech normalization across speaker, sex and accent variation is handled similarly by listeners of different language backgrounds

Gloria Pino Escobar[1,2], *Josephine Terry*[1,2], *Buddhamas Pralle Kriengwatana*[2,3], *Paola Escudero*[1,2]

[1]The MARCS Institute for Brain, Behaviour and Development,
Western Sydney University, Australia
[2]ARC Centre of Excellence for the Dynamics of Language, Canberra, Australia
[3]University of St Andrews, Scotland

{G.PinoEscobar, J.Terry, Paola.Escudero}@westernsydney.edu.au, bk50@st-andrews.ac.uk

Abstract

This study assessed the influence of language background in speech normalization by examining non-native vowel categorization across speaker, sex and accent variation. Mandarin-English bilinguals, Australian English bilinguals and monolinguals categorized /ɪ/ and /ɛ/ produced by a female Dutch speaker, and were then tested with the same vowels produced by speakers of the same or different sex and/or accent. Listeners categorized the vowels regardless of speaker and sex variation, but showed lower accuracy when vowels were produced by speakers of different accent or accent and sex. Findings suggest that listeners normalize speaker and sex variation automatically, while accent variation requires contextualization.

Index Terms: speech perception, normalization, vowel categorization, Mandarin, bilinguals, Go/No-Go task.

1. Introduction

Acoustic variation of phonologically identical sounds occurs across speakers of different sex, speakers of different accents, and even across same-sexed speakers of the same accent. Despite these differences, speakers of all languages are able to do away with this large variation and successfully communicate. For instance, listeners can disambiguate, recognize and discriminate sounds (phonemes, syllables or words) produced by speakers with different physical or personal characteristics [1] and learn to discriminate a sound even when carrying different acoustic properties [1, 2, 3].

Importantly, while acoustic variation across age and sex is largely attributed to differences in vocal-tract length [4, 5, 6], accent variation results from differences in speakers' language background due to geographical and/or socioeconomic factors [7, 8]. While accent variation has a sociolinguistic basis, speaker and sex variation are caused by both physiological and sociological factors [9]. As speaker and accent variation arise from different sources, this raises the possibility that listeners handle these two types of variation differently. Indeed, research on non-human animals suggests that speaker and sex normalization may be innate and pre-linguistic [10, 11]. For instance, Zebra finches exhibited accurate categorization of vowels produced by novel speakers of the same or different sex from the one they heard during training [10, 11], even when they were trained with only a single speaker and thus had no previous experience with normalizing speech across speakers [10]. In contrast, normalization of accent variation seems to require prior exposure to the specific accent or contextualization (e.g. awareness that a different accent is spoken). Specifically, accent variation initially obstructs speech comprehension, but after a period of exposure, the listener adapts to the accented sounds and succeeds at normalization [12].

Both behavioural and electrophysiological research has moved towards directly comparing listeners' handling of speaker/sex variation and accent variation to better understand whether these types of variation are handled differently, potentially with different mechanisms [9, 13, 14]. In one such study, electroencephalography (EEG) was used to measure participants' pre-attentive sensitivity to speaker, sex and accent variation [13]. Participants elicited larger mismatch negativity (i.e. the change-detection component of an event-related potential) when confronted with sex variation than when confronted with accent variation [13], which suggests that listeners are more pre-attentively sensitive to sex changes than to accent changes [13]. However, during a behavioural categorization task [9], Australian English (AusE) participants accurately categorized the Dutch vowels /ɪ/ and /ɛ/ across speaker and sex variation, but categorization performance declined when facing an accent or accent and sex change. Further investigations showed that AusE and Dutch listeners were able to successfully categorise Northern Dutch vowels /ɪ/ and /ɛ/ when confronted with speaker and sex changes but neither group were successful at categorising these vowels when they were produced by a speaker with a Flemish Dutch accent. This indicates that familiarity with the accent (in the case of Dutch speakers hearing Flemish-accented vowels) did not aid accent normalization. Interestingly, when the AusE participants were provided with feedback on their categorisation accuracy, they were able to successfully discriminate the Flemish-accented vowels. These results support the proposition that speaker and sex normalization occur automatically, while accent normalization requires contextualisation (e.g. feedback) [14]. Altogether, these results suggest that accent normalization and speaker/sex normalization are handled differently.

Although the aforementioned studies provide strong indicators towards disentangling the processes behind speech normalization, so far the primary subjects for these studies have been L1 English [9, 14] and L1 Dutch [14]. It is possible that speakers of varying linguistic backgrounds react differently when perceiving second-language (L2) vowels, as L1 vowel inventories have a direct impact on the way the L2 vowels are perceived [15, 16, 17]. If L1 experience affects perception, it may also affect the manner in which listeners normalize speech variation. Thus, it is possible that subjects with linguistic backgrounds other than English may perform differently from English speakers when required to normalize

Table of Contents for this manuscript

vowels from a non-native language. Currently, few studies have investigated differences and similarities in how native and non-native listeners normalize vowels.

The present study examines this hypothesis by directly comparing normalization of speaker, sex and accent variation in an unfamiliar language across participants with three different language backgrounds. We tested bilingual L1 Beijing Mandarin speakers' (Mandarin-BL) abilities to categorize the Dutch vowels /ɪ/ and /ɛ/ across speaker, sex and/or accent variation using a well-established behavioural Go/No-go categorization paradigm (see Methods) [9, 10, 11, 14]. We compared their performance to previously collected data from bilingual L1 Australian English (AusE-BL) and Monolingual Australian English (AusE-ML) speakers [9].

Isolated vowels were used as stimuli because they do not provide the listeners with contextual feedback (i.e. lexical, semantic or other) regarding the accuracy of their adaptation [14]. Complete words and natural speech would provide additional feedback during adaptation and therefore would hinder a neutral comparison between accent adaptation and speaker (and sex) adaptation as the latter seems to occur without this additional feedback [9, 10, 11, 14].

We tested Mandarin speakers for many reasons. First, Mandarin is the official language of China [18] and the most widely spoken language in the world [19]. It is also a L1 for the third largest group of overseas born Australians [20]. Furthermore, Mandarin is a tone language that differs from English in vowel inventory size and in its acoustic properties. While there are discrepancies regarding the number and classification of Mandarin vowels [15, 20], Mandarin has a smaller vowel inventory than AusE. For the purposes of this study we adopt the vowel classification by Zee and Lee [21] who consider Mandarin to have an inventory of seven monophthongs, /i, y, a, ə, ɚ,ɤ,u/. The Mandarin vowels are produced with four different tones, 1st Tone: High-Level, 2nd Tone: High-Rising, 3rd tone: Low- Dipping and 4th tone: High-Falling [20, 22, 23], which have the function of changing meaning [20]. In contrast, AusE has a larger vowel inventory than Mandarin that includes 12 monophthongs, /iː, ɪ, e, eː, ɜː, ɐ, ɐː, æ, o, ɔ, ʊ, ʉː/ [17].

The Second Language Linguistic Perception (L2LP) model posits that learners initially categorize sounds of the L2 to best match their L1 categories [16]. If the L1 contains fewer categories than the L2, the learner will face a difficult task in splitting or creating new categories when relying on duration or/and integrating spectral perception cues [16]. Given that Mandarin lacks the phonetic vowels /ɪ/ and /ɛ/ in their inventory, it was hypothesized that Mandarin speakers will categorize the vowels onto their closest Mandarin category. In this case Mandarin /i/ will be used to categorize Dutch /ɪ/; and Mandarin /i/ or /ɚ/ will be used to categorize Dutch /ɛ/. On the other hand, Australian English (AusE) has the same phonetic vowel /ɪ/, lacks /ɛ/ but has a similar vowel /e/. Having one counterpart for each of the vowel stimuli may make this task easier for the AusE-BL and AusE-ML listeners. Consequently, if language background affects categorization and normalization, it is predicted that both AusE groups will be more accurate than Mandarin listeners at categorizing the stimuli during a training phase.

In addition, L2 speakers are required to expand or adapt their L1 vowel inventory to create new mappings in order to perceive L2 categories [16]. L2 speakers may even activate a different auditory strategy not common to their L1 (e.g. relying on duration rather than on spectral cues [24]). With regard to bilingual AusE speakers (AusE-BL), possessing two languages could mean that the listeners' total vowel inventory size is the sum of the two individual languages, or at least equal to the language with the larger inventory if it contains a subset from the language with the smaller inventory. This situation could give the bilingual participants an advantage over monolinguals during a categorization task. Therefore, this study investigates whether L1 Mandarin speakers perform differently than L1 AusE speakers on normalizing speaker, sex and accent variation in vowel production and whether bilingualism or monolingualism has any significant effect on normalization.

We expect that if participants' ability to normalize speaker and sex variation is automatic and is not dependent on language background, then the three groups will maintain vowel discrimination performance when presented with a novel speaker of the same or different sex. Additionally, if language background does not play a role in accent normalization either, it is predicted that all three groups will have lower categorization accuracy when facing accent and accent+sex variation as they all lack experience with the new accent. Finally, it may be that the AusE-ML group would slightly underperform the AusE-BL groups, as bilinguals have prior experience adapting to new speech sounds.

2. Method

2.1. Participants

Participants were 30 adults (Mean age = 25.4 years, *SD* = 7.99, Range = 18-47 years), recruited from an Australian university and classified according to responses on a language background questionnaire. The Mandarin-BL group comprised 10 participants (8 females, Mean age = 27.7 years, *SD* = 6.82, Range = 22-46 years) who were native speakers of Mandarin Chinese and fluent in AusE, six of them additionally spoke another Chinese dialect or language. The AusE-BL group comprised 10 heterogeneous bilingual participants (4 females, Mean age = 21.4 years, *SD* = 6.02, Range = 18-38 years) who spoke AusE and one of the following languages: Arabic (3), Vietnamese (2), Egyptian (1), Macedonian (1), Serbian (1), Thai (1) and Hindi (1). The AusE-ML group comprised 10 native AusE monolinguals (9 females, Mean age = 27.1 years, *SD* = 9.83, Range = 19-47 years) who spoke no other languages. Data from the latter two groups was previously collected at the same Australian university and reported in [9]. None of the participants had previous experience with Dutch.

2.2. Stimuli and Procedure

Stimuli were natural isolated Northern and Flemish Dutch vowels /ɪ/ and /ɛ/, extracted from [s-Vowel-s] consonantal contexts spoken by male and female speakers [24]. Stimuli were presented within the Go/No-go task via headphones attached to a laptop computer running E-Prime (version 2) [9].

For the Go/No-Go behavioural categorization task, participants were required to respond to one vowel category (the "Go" vowel) by pressing spacebar within 2000 ms of the presentation of a stimulus, and to inhibit responses to the other vowel category (the "No-go" vowel), which required no action for 2000 ms after hearing a stimulus. Correct responses were reinforced with a smiley face and a pleasant bell sound, and participants were rewarded one point for each correct answer. False alarms and misses were penalized with a presentation of a sad face and a negative "punch" sound. No points were awarded for incorrect answers.

The task had three phases (*familiarization, training, testing)*. In the first phase (*familiarization*), easy to discriminate phonologically distinct words (*pon, deet*) were presented in order to familiarise participants with the task procedure. During the second phase (*training*), the Dutch vowels /ɪ/ and /ɛ/ produced by a female with a Northern Dutch accent were presented and feedback and points were provided. This phase aimed to train the participants to accurately categorize the two Dutch vowels. Table 1 describes the number of trials for each of the phases.

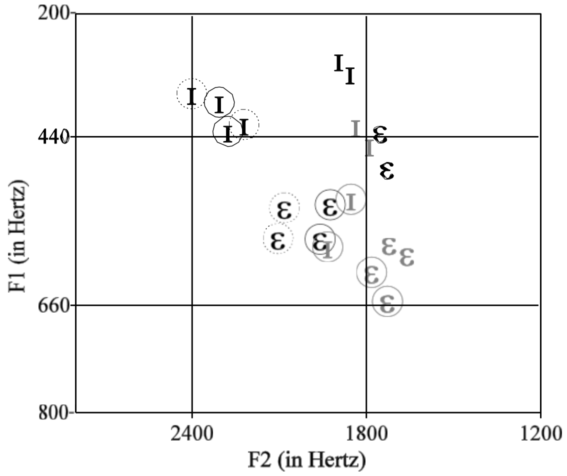


Figure 1: *Plot of Dutch vowels /ɪ/ and /ɛ/ used in the experiment. Vowels produced by a Northern Dutch female used in the training phase: Black circled. Novel Northern Dutch Female (speaker change): dotted circle. Northern Dutch male: Black without circle (sex change). Flemish*

Finally, in the third phase (*testing*), stimuli from training were presented intermixed with novel tokens of /ɪ/ and /ɛ/ produced by speakers who differed from the original speaker in the following ways: i) Speaker change (different female speaker with the same North Holland accent) ii) Sex change (male with North Holland accent), iii) Accent change (Female with Flemish accent) and iv) Accent+Sex change (male with Flemish accent). Figure 1 shows the F1 and F2 values of the stimuli in different conditions. No feedback was provided for responses to untrained stimuli and to an equal number of the trained stimuli (see Table 1). Only responses to the trained and novel stimuli without feedback (24 trials each) were analysed.

<table>
<tr><td colspan="3">Familiarization Phase
Pon/Deet: 20 trials</td></tr>
<tr><td colspan="3">Training Phase: Female Dutch
60 Trials</td></tr>
<tr><td rowspan="3"><b>Test Phase
120 trials
(100%)</b></td><td colspan="2">With Feedback: <i>Trained</i>: 72 trials
(60%)</td></tr>
<tr><td colspan="2"><b>Without Feedback: 48 trials (40%):</b></td></tr>
<tr><td><b>24
Trained Stimuli
(12 /ɪ/ and 12 /ɛ/)</b></td><td><b>24
Novel Stimuli
(12 /ɪ/ and 12 /ɛ/)</b></td></tr>
</table>

Table 1: *Number of trials per phase. 48 trials without feedback were used for analysis.*

A between-subjects design was employed with each participant tested in only two of the conditions: i) speaker and sex (Mandarin-BL = 5, AusE-BL = 5 and AusE-ML = 5) or ii) accent and accent+sex (Mandarin-BL = 5, AusE-BL = 5 and AusE-ML = 5). The order in which participants performed each condition was counterbalanced in each language group.

3. Results

For each participant, a difference score was computed by subtracting test accuracy (i.e. to trials without feedback) from training accuracy (% correct). These difference scores were compared in an ANOVA with speaker-change condition (i.e. *speaker, sex, accent and accent+sex*) and language background as between-subjects factors. This revealed a main effect of Speaker-change condition, $F(3, 48) = 16.77$, $p <.01$, partial $\eta 2 = .51$. Pairwise comparisons revealed that the accuracy difference between the test and training phases was greater in the accent and accent+sex conditions compared to the speaker and sex conditions, *ps* < .01, which indicates that listeners in these conditions performed with relatively lower accuracy when presented with a novel speaker of a different accent or accent and sex. There was no difference in accuracy scores between the speaker and sex conditions, nor between the accent and accent+gender conditions, *ps* > .46. There was also no main effect of language background and no interaction with speaker-condition, *ps* > .67 (See Figure 2).

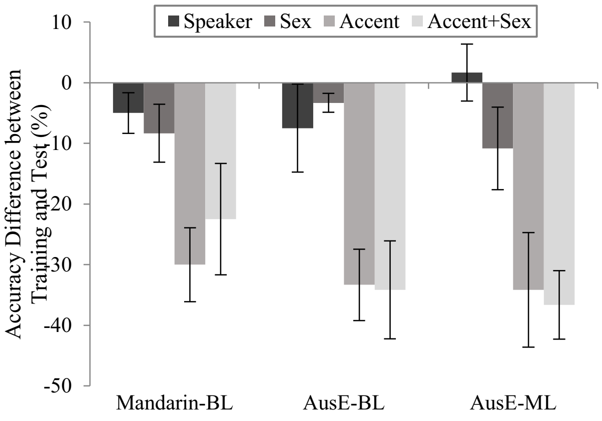


Figure 2. *Accuracy difference scores: negative scores indicate a decrease in accuracy at the test phase (compared to the training phase).*

4. Discussion

Results show that listeners from different language backgrounds who are naive to Dutch vowels and to Northern Dutch and Flemish accents are able to normalize speaker and sex variation. When facing speaker or sex variation, the three groups maintained the same level of accuracy that was evident with trained vowels. In other words, Mandarin-BL, AusE-BL and AusE-ML listeners were able to categorize the vowels with similar accuracy when produced by an unfamiliar Northern Dutch female and an unfamiliar Northern Dutch male. Results were different when facing accent and accent+sex variation. Across all language groups, categorization accuracy was poor when the vowels were produced by a female with a Flemish Dutch accent and by a male with Flemish Dutch accent. These results support the proposition that listeners have an intrinsic ability to normalize variations in speaker and sex but not accent, which suggests that listeners may need additional exposure to the accent-varied stimuli in order to categorize it accurately [12,14].

Findings were consistent across the three languages groups. Predictions that AusE-BL and AusE-ML would outperform Mandarin-BL listeners' performance in Speaker and Sex conditions, due to similarities in vowel inventories between AusE and Dutch, were not confirmed. Indeed, results in categorization accuracy of the three language groups do not show any significant differences. Moreover, Mandarin listeners were able to categorize Dutch vowels /ɪ/ and /ɛ/ accurately in the training phase. This could have possibly occurred because the inventory of Mandarin vowels, despite having only seven phonetic vowels, has four tones variations. That is, considering that the four tones have different acoustic properties, this may mean that Mandarin listeners have a richer acoustic space in their vowel inventory, making them more sensitive to vowel contrasts than the AusE participants. Finally, no difference was found between the AusE-BL and AusE-ML groups. This suggests that our study was not able to find an effect of bilingualism or linguistic background on the normalization of isolated vowels, which may be due to the small number of participants included in each group. Further research comparing the same three groups with a larger listener sample should be conducted to confirm the current findings.

In conclusion, results of the present study support earlier findings that listeners may use an automatic or innate mechanism to normalize speaker and sex variation [9, 13, 14]. Further, these results give preliminary evidence suggesting that the difficulties in normalizing across accent variation may not have a language-specific basis. Besides including a larger sample of participants, future research could also investigate how pre-exposure to an accent affects perception by listeners of different language backgrounds.

5. Acknowledgements

We thank Samra Alispahic and Rozmin Dadwani for their assistance with data collection. This study was funded by the MARCS Institute for Brain, Behaviour & Development, the Australian Research Council (ARC) Discovery Grant to Paola Escudero [DP 130102181] and the ARC Centre of Excellence for the Dynamics of Language

6. References

[1] Traunmüller, H., "Articulatory and perceptual factors controlling the age-and sex-conditioned variability in formant frequencies of vowels", Speech Communication, Vol 3, no. 1, pp. 49-61, 1984.

[2] Fant, G., "Non-uniform vowel normalization". Speech Transmission Laboratory Quart. Progress and Status Report, Vol 16, nos. 2-3, pp. 1-19, 1975.

[3] Clarke, C.M. and Garrett, M.F., "Rapid adaptation to foreign-accented English". The J. of the Acoustical Soc. of America, Vol 116, no.6, pp. 3647-3658, 2004.

[4] Fitch, W. T. and Giedd, J., Morphology and development of the human vocal tract: A study using magnetic resonance imaging. J. of the Acoustical Soc. of America, Vol. 106, pp. 1511-1522, 1999.

[5] Huber, J. E., Stathopoulos, E. T., Curione, G. M. Ash, T. A. and Johnson, K., Formants of children, women, and men: The effects of vocal intensity variation. J. of the Acoustical Soc. of America, Vol.106, pp. 1532-1542, 1999.

[6] Monahan, P. and Idsardi, W. Auditory sensitivity to formant ratios: Toward an account of vowel normalisation. Language and Cognitive Processes, Vol. 25, no.6, pp. 808 -839, 2010.

[7] Adank, P., Noordzij, M.L. and Hagoort, P. "The role of planum temporale in processing accent variation in spoken language comprehension" Human brain mapping, Vol 33, no. 2, pp. 360-372, 2012.

[8] Evans, B. G., and Iverson, P "Vowel normalization for accent: An investigation of best exemplar locations in northern and southern British English sentences", The J. of the Acoustical Soc. of America, Vol. 115, pp. 352-361, 2004.

[9] Kriengwatana, B., Escudero, P. and Terry, J., "Listeners cope with speaker and accent variation differently: Evidence from the Go/No-go task". 15th Australasian Intl Conf. on Speech Sci. and Technology, 2014.

[10] Kriengwatana, B., Escudero, P., Kerkhoven, A. and ten Cate, C. "A general auditory bias for disregarding inter-speaker speech variability: Evidence in humans and songbirds", Front. Psychol. Vol. 6, pp. 1234, 2015.

[11] Ohms, V. R., Escudero, P., Lammers, K. and ten Cate, C., "Zebra finches and Dutch adults exhibit the same cue weighting bias in vowel perception", Animal cognition, Vol. 15, no.2, pp. 155-161, 2012.

[12] A. Cristia, A. Seidl, C. Vaughn, R. Schmale, A. Bradlow and C.Floccia, "Linguistic processing of accented speech across the lifespan". Frontiers in psychology, Vol. 3, 2012.

[13] Dadwani, R., Peter, V., Chladkova, K., Geambesu, A., Escudero, P., "Adult listeners' processing of indexical versus linguistic differences in a pre-attentive discrimination paradigm", Proceedings of the 18th International Congress of Phonetic Sciences (ISBN 9780852619414), 2015.

[14] Kriengwatana, B., Terry, J., Chládková, K., & Escudero, P. (2016). Speaker and Accent Variation Are Handled Differently: Evidence in Native and Non-Native Listeners. PloS one, 11(6), e0156870.

[15] J. E. Flege, O.-S. Bohn and S. Jang "Effects of experience on non- native speakers' production and perception of English vowels," J. of Phonetics, Vol. 25, no 4, pp. 437-370, 1997.

[16] Escudero, P., Linguistic Perception and Second Language Acquisition. "Explaining the attainment of optimal phonological categorization" Ph.D. dissertation, Utrecht Univ. LOT Dissertation Series Vol. 113, 2005.

[17] Elvin, J., Escudero, P. & Vasiliev, P., 2014, "Spanish is better than English for discriminating Portuguese vowels: acoustic similarity versus vowel inventory size, Frontiers in psychology, 2014.

[18] Chen, Y., Robb, M., Gilbert, H. and Lerman, J., "Vowel production by Mandarin speakers of English" Clinical Linguistics & Phonetics, Vol. 15, no. 6, pp. 427-440, 2001.

[19] Singh, L. and Foong, J., "Influences of lexical tone and pitch on word recognition in bilingual infants". Cognition, Vol. 124, no. 2, pp. 128-142, 2012.

[20] Australian Bureau of Statistics, Australian population by country of birth. Available FTP http://www.abs.gov.au/ausstats/abs@.nsf/Lookup/3412.0Chapter 12011-12%20and%202012-13, 2014.

[21] Zee, E., & Lee, W. S. (2001). "An acoustical analysis of the vowels in Beijing Mandarin". In EUROSPEECH Proc., pp. 643-646, 2001.

[22] Chao, Y. R., A grammar of spoken Chinese. Univ of California Press, 1968

[23] Howie, J. M., Acoustical studies of Mandarin vowels and tones, No. 6, Cambridge Univ. Press, 1976.

[24] Escudero, P. and Boersma, P., "Bridging the gap between L2 speech perception research and phonological theory," Studies in Second Language Acquisition, Vol 26, no. 04, pp. 551-585, 2004.

[25] Adank, P., Van Hout, R. and Smits, R., "An acoustic description of the vowels of northern and southern standard Dutch," J. Acoust. Soc. Am. 116, pp 1729–1738, 2004.

Cross-accent word recognition is affected by perceptual assimilation

Sarah M. Wright[1], *Mark D. Lathouwers*[2], *Catherine T. Best*[1,3,4], *Michael D. Tyler*[1,2]

[1]The MARCS Institute, Western Sydney University, Australia
[2]School of Social Sciences & Psychology, Western Sydney University, Australia
[3]School of Humanities & Communication Arts, Western Sydney University, Australia
[4]Haskins Laboratories, New Haven, Connecticut, USA

sarah.wright@westernsydney.edu.au

Abstract

A single-item shadowing task was conducted to determine how identification of London-accented words by Australian listeners is affected by perceptual assimilation. This was evaluated in conjunction with two other well-established effects on word recognition: word frequency and talker variability. The results replicate frequency and talker variability effects, support the hypothesis that talker and accent normalisation operate at different processing stages, and show that words with nativelike assimilation of all phonemes are identified more accurately than those with category goodness or category shifting assimilation. Results are evaluated in view of episodic theories of lexical access.

Index Terms: cross-accent speech perception, perceptual assimilation, shadowing, word frequency, talker variability

1. Introduction

Variation in speech results from a variety of factors including individual differences in speaker gender, vocal tract characteristics, and speaker origin [1]. Despite this, the average person usually quite easily understands words across their many varying forms without any conscious awareness of adjusting for talker or token variability.

Accented speech is commonly encountered, and may affect comprehension and speed of processing, at least initially. For example, although initially disturbed, comprehension of an unfamiliar accent returns to nativelike levels within one minute of listening [2]. Additionally, [2] suggests that comprehension and speed of processing may vary depending on how 'thick' the accent is perceived to be. That is, accents can be ranked on a perceptual scale as a function of their distance from the native accent with foreign accents being more accented compared to regional accents of the native language. This suggestion is supported by their finding of a 100-150 ms delay in word identification in a foreign accent compared to the native accent, whereas [3] found only a 30 ms delay in word identification by French participants listening to an unfamiliar French accent.

However, even anecdotally, there are likely to be situations where the perceived strength of an accent is not related to whether it is a foreign or regional accent to the listener. For example, an Australian English speaker may perceive a Dutch accent as less accented, and therefore more comprehensible, compared to a regional accent of English such as Glaswegian.

The Perceptual Assimilation Model (PAM) [4-6] offers an alternative way of using phonetic and phonological differences to categorise accent strength in an unfamiliar accent compared to a perceiver's own accent. Although PAM was designed to account for perceptual assimilation of non-native phones, its principles can also be applied to cross-accent perception [7]. Phonemes in the unfamiliar accent that are perceived as good exemplars of the same phonological category in the native accent are assimilated to the native accent as *nativelike* (NL). A *category goodness* (CG) assimilation occurs when a non-native phoneme is perceived as an acceptable but not ideal exemplar of the same category in the native accent. It should be identified as the same phonological category but deviate in goodness of fit. Finally when a phoneme in the unfamiliar accent is sufficiently disparate from the native accent to be perceived as a different native phonological category, it is a *category shifting* (CS) assimilation. This model proposes that that individual words in an unfamiliar accent may be perceived as more or less accented depending on which phonemes the words contain and how those phonemes are assimilated to native phonological categories. Therefore, the main aim of the present study is to investigate how assimilation type affects cross-accent word recognition. We hypothesise that word recognition will be most accurate and efficient for NL assimilations, followed by CG, then CS types.

Speech from multiple talkers introduces variability that is less systematic and narrower in scope than variability across accents [5]. According to episodic theories of lexical access, variability due to talkers and accents is stored in long-term memory and is used to facilitate word recognition as opposed to the pre-lexical identification of phonemes as proposed by abstractionist theories [5]. Talker variability effects are well established. For example, word identification in a single-item shadowing task was slower and less accurate when words were produced by 15 different talkers, compared to those produced by a single talker [8], indicating that talker variability may not be removed prior to lexical access. In addition, high-frequency words were identified more accurately (but not more quickly) than low-frequency words. Thus, both word frequency and amount of talker variability affect the time course and/or accuracy of lexical access in the native accent, but little is known about how cross-accent assimilation is modulated by the presence of these other types of variation. To assess this, word recognition across accents was tested. Australian participants repeated aloud [following 8] high- and low-frequency words spoken in either their native Australian accent (AusE) or the much less familiar Southeast London accent, by either one or multiple speakers.

2. Method

2.1. Participants

Forty-eight introductory psychology students (41 females) at Western Sydney University participated for course credit.

Participants were aged between 18 and 35 years (*M* = 21.15, *SD* = 3.57), had all been exposed to AusE from birth, and reported no hearing or speech disorders at the time of testing. An additional 12 participants were tested but their data were discarded for incomplete responses (*n* = 1), previous exposure to a UK accent (*n* = 4), not being exposed to AusE from birth (*n* = 4), being outside of the 18-35 age range (*n* = 1), and equipment malfunction *(n* = 2*).*

2.2. Research Design

The experiment was a 2 x 2 x (2 x 2 x 3) mixed-design with reaction time (RT) and accuracy as dependent variables. Number of talkers (single vs. multiple) and accent (AusE vs. London) were manipulated between subjects, and word frequency (low vs. high), number of syllables (1 vs. 2), and London word assimilation type (NL vs. CG vs. CS) were manipulated within subjects. Participants were randomly assigned to one of the four between-subject cells (*n* = 12).

2.3. Stimulus Materials

The target stimuli consisted of 132 words selected from an existing cross-accent spoken word corpus. These words were selected based on assimilation type, frequency and number of syllables. See Table 1 for item totals across these variables.

The cross-accent assimilation types were determined based on published phonetic descriptions of the London and Australian accents. For example, all the phonemes in London-accented *baby* ([bæɪbi]) should be assimilated as good exemplars of the same native phonological categories (an NL assimilation). In the London-accented word *note* ([nəʉʔ]), all phonemes are assimilated as good exemplars of the same native categories, except for /ʔ/, which is assimilated as a poor exemplar of the native /t/ category (a CG assimilation). For London-accented *thorny* ([fo:ni]), all phonemes are assimilated as good exemplars of their native phonological categories except for the initial consonant which is assimilated to the native /f/ category instead of the intended /θ/ target (a CS assimilation). High-frequency words had a frequency of 50 or more per million (*M* = 214.5, *SD* = 127.1) and low frequency items had less than 10 occurrences per million (*M* = 2.8, *SD* = 2.4).

Table 1: *Mean duration (ms), mean frequency per million, and number of items by number of syllables, assimilation type, frequency, and accent.*

Syllables	Assimilation	Frequency	Accent	M_{dur}	M_{freq}	n_{items}
1	Nativelike	High	AusE	590	243	12
			London	549		
		Low	AusE	568	4	12
			London	529		
	Category Goodness	High	AusE	575	167	12
			London	470		
		Low	AusE	581	3	12
			London	524		
	Category Shifting	High	AusE	582	256	8
			London	532		
		Low	AusE	586	3	8
			London	509		
2	Nativelike	High	AusE	615	216	12
			London	632		
		Low	AusE	653	2	12
			London	660		
	Category Goodness	High	AusE	645	196	12
			London	640		
		Low	AusE	644	3	12
			London	642		
	Category Shifting	High	AusE	573	210	10
			London	585		
		Low	AusE	631	2	10
			London	657		

There were six stimulus lists per between-subjects condition, and each was used twice across the 12 participants in each group. In the single-talker condition, there was one list for each of the six speakers. In the multiple talker condition, each speaker's words were distributed across the six lists using a Latin square design, such that each token was presented an equal number of times across the single- and multiple-talker groups. The presentation was blocked by assimilation type with high- and low-frequency words presented randomly within each. NL and CG blocks were presented first in counterbalanced order. The CS block was always presented last, as the smaller number of items meant it might need to be excluded at a later stage.

2.4. Apparatus

The experiment was controlled using PsyScope X [9] running on a MacBook with an Edirol sound card. Participants listened to the stimuli through headphones and responded into a headset microphone (Beyerdynamic DT290). Trials were progressed using a voice key, and then reaction time was calculated manually from the recorded waveform, from the onset of the stimulus to the onset of the participant's response.

2.5. Procedure

Each trial began with the word *ready* displayed on the screen followed by the auditory presentation of the test item. Participants were instructed simply to repeat the word that they heard as quickly and as accurately as possible. If they took longer than 3.5 s to respond they were instructed to respond more quickly. There were no breaks between blocks. At the completion of the shadowing task, participants completed a language background questionnaire.

3. Results

RTs were analysed for correct responses only. Trials were scored as incorrect if a response was a word other than the test item, except for plurality errors (e.g., *papers* instead of *paper*; 8.33% of all trials). Responses that timed out were also considered incorrect (0.28% of all trials). RT and accuracy data were analysed using analysis of variance with planned contrasts (see 2.2 for the research design). The assimilation type variable was analysed using three non-orthogonal contrasts: 1) NL vs. CG, 2) NL vs. CS, and 3) CG vs. CS. An alpha level of .019 was used to adjust for using multiple contrasts [see 10]. For brevity only main effects and significant interactions involving the accent variable are reported below.

3.1. Reaction Time

There were four significant main effects for reaction time. Reaction times were shorter when words were presented from a single speaker (812 ms) compared to multiple speakers (918 ms), $F(1, 44) = 8.55$, $p = .005$, $\eta_p^2 = .16$, and for words in their native accent (788 ms) compared to the London accent (942 ms), $F(1, 44) = 18.00$, $p < .001$, $\eta_p^2 = .29$. High-frequency words (819 ms) were responded to more quickly than low-frequency words (911 ms), $F(1, 44) = 233.83$, $p < .001$, $\eta_p^2 = .84$, and participants responded more quickly to 1-syllable (840 ms) than 2-syllable (890 ms) words, $F(1, 44) = 125.38$, $p < .001$, $\eta_p^2 = .74$. There were no significant effects of assimilation type on RT.

These main effects were moderated by two significant interactions. An accent × frequency interaction, $F(1, 44) = 44.57$, $p < .001$, $\eta_p^2 = .50$, indicates that the mean difference between low- and high-frequency words was significantly greater in the London accent ($M_{\text{diff}} = 133$ ms) compared to the Australian accent ($M_{\text{diff}} = 52$ ms). An accent × syllable interaction, $F(1, 44) = 26.96$, $p < .001$, $\eta_p^2 = .38$, indicates that the mean difference between 1- and 2-syllable words was significantly greater in the London accent ($M_{\text{diff}} = 73$ ms) compared to the Australian accent ($M_{\text{diff}} = 27$ ms).

3.2. Accuracy

The accuracy results are presented in Figure 1. They are collapsed across the number of talkers variable because it did not interact significantly with any other factors. Overall mean accuracy was higher for the Australian accent (96%) compared to the London accent (87%), $F(1, 44) = 134.22$, $p < .001$, $\eta_p^2 = .75$, for the single-talker (92%) than the multiple-talker condition (90%), $F(1, 44) = 11.24$, $p = .002$, $\eta_p^2 = .20$, for high-frequency (93%) than low-frequency items (89%), $F(1, 44) = 66.78$, $p < .001$, $\eta_p^2 = .60$, and for 2-syllable words (95%) than 1-syllable words (87%), $F(1, 44) = 100.30$, $p < .001$, $\eta_p^2 = .70$. The three non-orthogonal contrasts on the assimilation type variable were all significant: NL vs. CG: $F(1, 44) = 27.50$, $p < .001$, $\eta_p^2 = .38$; NL vs. CS: $F(1, 44) = 46.93$, $p < .001$, $\eta_p^2 = .52$; and, CG vs. CS: $F(1, 44) = 7.33$, $p = .01$, $\eta_p^2 = .14$. These results show that accuracy was significantly different across all levels of assimilation type with NL items having the highest accuracy (95%) followed by CG (90%) and then CS (88%). Note, however, that this is collapsed across the accent factor, and that the assimilation types are only relevant for the words when they are spoken in a London accent.

The main effects were moderated by four significant two-way interactions and a three-way interaction. The first two-way interaction was between accent and frequency, $F(1, 44) = 7.70$, $p = .008$, $\eta_p^2 = .14$. As can be seen in Figure 1, the difference in accuracy between high- and low-frequency words was greater for the London accent compared to the Australian accent. The interaction between accent and syllable, $F(1, 44) = 24.09$, $p < .001$, $\eta_p^2 = .35$, indicates that the difference in accuracy for words spoken in AusE and London accents is greater for 1-syllable than 2-syllable words. Additionally, these two effects were moderated by a three-way interaction of accent, frequency and syllable, $F(1, 44) = 6.91$, $p = .012$, $\eta_p^2 = .14$. Figure 1 shows that the greater effect of low frequency than high frequency words on accuracy in the London versus AusE accent was more pronounced for the 1-syllable than the 2-syllable words.

Finally, two of the assimilation type contrasts interacted significantly with accent: NL vs. CG, $F(1, 44) = 33.80$, $p < .001$, $\eta_p^2 = .43$; and NL vs. CS, $F(1, 44) = 26.16$, $p < .001$, $\eta_p^2 = .37$. These interactions show that both CG and CS words were recognised less accurately than NL words in the London accent compared to the AusE. However, the absence of a significant interaction between accent and the CG vs. CS contrast indicates that the mean difference in accuracy for the CG vs. CS items in AusE (2%) is not significantly different from the mean difference of the accuracy for the same items in the London accent (3%). This suggests that the significant main effect contrast between CG words and CS words is due to inherent properties of the words themselves rather than to pronunciation differences when those words are spoken in AusE versus London accents.

4. Discussion

Consistent with previous research, participants were less accurate and took longer to recognise words in a non-native regional accent than in their own accent. This is consistent with episodic theories of lexical access that suggest that items falling to the edges of categories (pronunciations that deviate

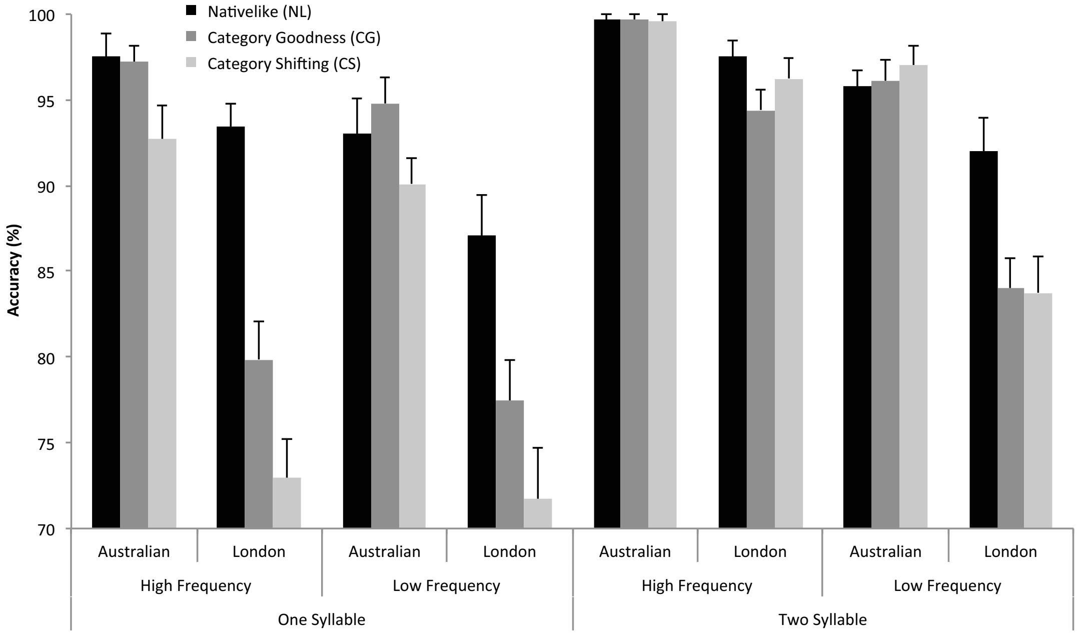


Figure 1: *Accuracy data collapsed across number of talkers, separated by accent, word frequency, number or syllables, and assimilation type. Error bars represent standard error of the mean.*

from the native accent) will take longer to be identified than items that are more typically encountered as in the native accent. Crucially, however, the effect of accent was moderated by assimilation type. Word recognition in the London accent was perturbed less for words that were assimilated as nativelike than for words that are phonetically and/or phonologically different from the native pronunciation (CG and CS assimilations). Contrary to our hypothesis, however, we observed no relative difference in accuracy or reaction time between category-goodness and category-shifting assimilations. However, as there were a smaller number of CS items compared to CG and NL items, due to the constraints of the existing corpus, it is possible that the number of observations was insufficient to provide a reliable estimate of the accuracy difference between CG and CS items. Future research should investigate this with a larger stimulus sample. Nevertheless, our results show a clear effect of perceptual assimilation on cross-accent word recognition between the NL items and the CG and CS items. This indicates that not all accent differences are equal and that accuracy is likely to be higher for items that are assimilated as good exemplars of native phonological categories, than phonetically and/or phonologically mismatching exemplars. We suggest, therefore, that perceptual assimilation may also play a role in recognition of foreign-accented speech.

The main effect of speaker number on reaction time replicates the well-established effect of number of speakers on word recognition [8], consistent with episodic models. However, the lack of an interaction between number of speakers and accent for either the accuracy or RT data suggests that the mechanisms for normalising across multiple speakers and normalising across accents may be operating at different stages of processing. For example, it has been suggested that speaker normalisation may occur at a pre-lexical stage and accent normalisation at a lexical stage [11], the latter being more consistent with abstractionist models.

The main effect of frequency as well as the frequency by accent interaction for RT also provides support for episodic theories of lexical access. That is, the robust frequency effects found in previous studies (e.g., [8]) have been replicated. It is also apparent that this frequency effect is moderated by accent. This is likely to correspond to the number of typical (native) versus deviant (other-accent) episodes held in memory, and that this number is reduced more so for low-frequency words compared to high-frequency words. This finding is also supported by the main effect of frequency, and the accent by frequency interaction for the accuracy data.

The main effect of syllable is unremarkable as it is expected that one-syllable words would be identified more quickly than two syllable words. The accent by syllable interaction for the RT data should be interpreted with caution in this case, as stimulus length was not controlled across accents. However, the accent by syllable interaction for the accuracy data is an interesting finding. Although it is expected that one syllable words will have lower accuracy than two syllable words, as shorter words tend to have more lexical neighbours [12], this effect appears to be exacerbated by the phonetic ambiguity introduced by a non-native regional accent. Additionally, the three-way interaction among syllable, accent and frequency for the accuracy data suggests that the ambiguity introduced by the non-native accent cancels out the effects of any frequency benefit for one-syllable words, which are more easily confused with each other due to their larger lexical neighbourhoods, as compared to two-syllable words.

Finally, on a methodological note, the shadowing task is useful because it measures performance at a point after word recognition has occurred. However, the obtained results, particularly RT, may have been affected by the reliance of this task on speech planning and production. Future research should employ a purely perceptual task such as lexical decision or a visual world word recognition paradigm to tease out this issue.

5. Conclusions

This study replicated the well-established effects of talker variability and word frequency on speed and accuracy of lexical access in a single-item shadowing task. It provides support for the idea that the mechanisms for normalising across accents and multiple talkers may operate at different stages of processing. Finally, it produces a new finding that not all accent differences impede lexical access equally. Rather, cross-accent word recognition is moderated by assimilation type: NL items in the London accent are identified more accurately than CG or CS items.

6. Acknowledgements

This research was funded by ARC grant DP120104596.

7. References

[1] D.B. Pisoni & S.V. Levi, "Some observations on representations and representational specificity in speech perception and spoken word recognition," in *The Oxford handbook of psycholinguistics*, M.G. Gaskell, Ed. Oxford, UK: Oxford University Press, 2007, pp 3-18.

[2] C.M. Clarke & M.F. Garrett, "Rapid adaptation to non-native speech," *J. Acoust. Soc. Amer.*, vol. 116, no. 6, pp. 3647-3658, 2004.

[3] C. Floccia *et al.*, "Does a regional accent perturb speech processing?," *J. Exp. Psych.: Human Per. and Perf.*, vol. 32, no. 5, pp. 1276-1293, 2006.

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

[5] C.T. Best, "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.

[6] C.T. Best & M.D. Tyler, "Nonnative and second-language speech perception: Commonalities and complementarities," in *Seconds language learning: The role of language experience in speech perception and production*, M.J. Munro & O.S. Bohn, Eds. Amsterdam, The Netherlands: John Benjamins, 2007, pp. 13-34.

[7] C.T. Best *et al.*, "From Newcastle MOUTH to Aussie ears: Australians' perceptual assimilation and adaptation for Newcastle UK vowels," *Interspeech*, Dresden, Germany, 2015.

[8] J.W. Mullennix, D.B. Pisoni, & C.S. Martin, "Some effects of talker variability on spoken word recognition," *J. Acoust. Soc. Amer.*, vol. 85, no. 1, pp. 365-378, 1989.

[9] L.L. Bonatti, PsyScope X (Build 57). Trieste, Italy: SISSA, 2010.

[10] M.A. Betz & J.R. Levin, "Coherent analysis-of-variance hypothesis-testing strategies: A general model," *J. Edu. Stat.*, vol. 7, no. 3, pp. 193-206, 1982.

[11] B. Kriengwatana *et al.*, "Speaker and accent variation are handled differently: Evidence in native and non-native listeners," *PLoS ONE*, vol. 11, no. 6, 2015.

[12] D.B. Pisoni *et al.*, "Speech perception, word recognition and the structure of the lexicon," *Speech Comm.*, vol. 4, no. 1-3, pp. 75-95, 1985.

Comparison of Speech Enhancement Algorithms for Forensic Applications

Ahmed Kamil Hasan, David Dean, Bouchra Senadji and Vinod Chandran

School of Electrical Engineering and Computer Science , Queensland University of Technology

ahmedkamilhasan.alali@hdr.qut.edu.au, ddean@ieee.org, {b.senadji, v.chandran}@qut.edu.au

Abstract

Speech enhancement algorithms play an essential role in forensic applications, and enhanced speech signals can be used in court as evidence in criminal cases. This paper compares the performance of single channel (spectral subtraction and level dependent wavelet threshold techniques) and multiple channel (independent component analysis or ICA) speech enhancement algorithms to remove real environmental noise from noisy audio recording signals. Experimental results demonstrate that ICA achieves a significant improvement in average signal to noise ratio (SNR) enhancement compared to single channel speech enhancement algorithms, when 100 sentences from a forensic voice comparison database were corrupted with a car, street and factory noise at input SNR (-10 to 10 dB).

Index Terms: speech enhancement, independent component analysis, spectral subtraction, wavelet threshold technique

1. Introduction

Speech recordings obtained in the context of law enforcement agencies are often degraded by various types of real environmental noise. The police agencies often use hidden microphones to record the speech from the criminal in public places. Such forensic audio recordings may be far away from the hidden microphones and these recordings are often corrupted by car, street or machinery (factory) noise. It is difficult to directly use these recordings in court as a part of evidence in criminal cases, because their intelligibility is poor. Therefore, speech enhancement algorithms in real-life casework may be more complicated than those in theoretical research. Choosing the most reliable method for speech enhancement algorithm under these conditions play an important role in forensic applications. The enhanced speech signal can be used to eventually establish or confirm the identity of the criminal [1].

Speech enhancement algorithms can be divided into single channel and multiple channel algorithms depending on the number of the microphones that are used for collecting the noisy speech signal. Various algorithms for single channel speech enhancement, such as spectral subtraction [2] and wavelet threshold techniques [3] have been proposed in the last few decades, but these methods do not achieve great improvements in speech quality when the speech signal is corrupted with real environmental noise.

The spectral subtraction algorithm [2] is based on subtracting the estimated spectrum of the noise signal from the spectrum of the noisy speech signal. Since the spectrum of real environmental noise and speech signal are not uniformly distributed over the whole frequency bands, the musical noise will appear in the enhanced speech signal. This noise will lead to reduction in the quality of the denoised speech signal [4]. Wavelet denoising techniques are widely used to suppress noise from noisy speech signals [3] [4]. Noise is removed by applying an appropriate threshold to the wavelet cofficients for high frequency bands (detailed coefficients). This is based on the assumption that detail coefficients below significant energy levels arises from background noise rather than speech [3]. Wavelet threshold techniques fail to suppress noise in high SNR cases [5]. Colored noise is a non-stationary signal and the distribution of colored noise is spread non-uniformly over different frequency sub bands [4]. Such noise can have significant energy in the wavelet coefficients for low frequency band (approximation) or detail wavelet coefficients. If the power spectral density of the colored noise is concentrated at low frequency sub bands, a threshold applied to high frequency components of the noisy speech signals will not eliminate the low frequency components of the noise and will lead to a poor signal to noise ratio at the output.

Multichannel speech enhancement algorithms can be used to suppress and improve the quality of the speech signal under noisy conditions [6] . Independent component analysis (ICA) is widely used in multi channel speech enhancement and it is used to separate the speech from the noise by transforming the noisy speech signals into components which are statistically independent [7]. The principle of estimating independent components is based on maximizing the non-Gaussian distribution of one independent component [8]. The difference between a Gaussian distribution and the distribution of the independent component is measured using a contrast parameter, such as kurtosis, which is maximized by the ICA algorithm [8].

Single and multiple speech enhancement algorithms were also used to suppress the noise from noisy forensic recording signal in the existing literature review. Single channel speech enhancement was used with dynamic time warping and wavelet packet threshold techniques to suppress co-talker interference noise from forensic audio recordings in [9]. Multichannel speech enhancement was used to remove co-talker noise from the noisy forensic recording by using the delay and sum beamforming algorithm in [6]. Spectral subtraction was used to remove colored noise from mixed speech signals and convolutive ICA was used to separate one speaker from another in [1] to improve the performance of speaker identification. The original contribution of this research is to investigate the performance of ICA to suppress real environmental noise from short utterance of noisy forensic recordings, and to compare this performance with single speech enhancement algorithms under such conditions.

2. Model of ICA

Let the speech and noise signals emitted from N sources be represented as $s(t) = \{s_1(t), s_2(t), ..., s_N(t)\}$. The noisy speech signals can be recorded instantaneously by using M microphones in a street for forensic applications and be expressed as $x(t) = \{x_1(t), x_2(t), \cdots, x_M(t)\}$. Instantaneous ICA can be defined as a linear transformation of noisy speech signals into components which are statistically independent, and can be represented as [8]

$$x = As \tag{1}$$

where A is an unknown mixing matrix

The goal of ICA is to estimate the original sources from the mixed signals. The estimates of speech and noise signals $(\hat{s})$ can be represented by the following equation:

$$\hat{s} = Wx \tag{2}$$

where W is the unmixing matrix which equals the inverse of the mixing matrix A when the matrix is square.

In this paper, we use two sources (speech and noise) and two microphones to record the noisy speech signals$(M = N = 2)$. Therefore, the mixing and unmixing matrices are square and they have a size of 2×2.

2.1. Fast ICA Algorithm

The procedure for a fast ICA algorithm for one unit can be illustrated by the following steps [8] :

1. Remove the mean value from the noisy signal and center its distribution.
2. Whiten the noisy speech signal (x) to get (x_w) by using eigenvalue decomposition of the covariance of the noisy speech signal.

   $$x_w = VD^{-1/2}V^T x \tag{3}$$

   where V is the eigenvector matrix of the covariance of the noisy speech signal, and $D^{-1/2}$ is the inverse square root diagonal of the eigenvalue matrix.
3. Choose an initial vector of unmixing matrix W.
4. Estimate a row vector of unmixing matrix

   $$w^+ = E\{x_w\, g(w^T x_w)\} - E\{g'(w^T x_w)\}w \tag{4}$$

   where w^+ is the new value of the row vector of the unmixing matrix, E is the sample mean, g and g' are the first and the second derivatives of the contrast function respectively.
5. Normalize the row vector of w^+

   $$w^* = \frac{w^+}{\|w^+\|} \tag{5}$$

   where w^* is the normalization of the new row vector of the unmixing matrix.
6. Insert $w = w^*$ in step 4 and repeat the procedure until there is convergence.

The criterion of convergence is that the direction of previous and new values of w must be in the same direction, i.e. the dot product of these w points is almost equal to one.

This algorithm is based on separating one non-Gaussian component each time under the assumption that the sum of the others has a Gaussian distribution. It is necessary to prevent different row vectors of w from converging to the same maxima and this can be performed by using a deflation decorrelation of the output $w_1^T x, w_2^T x, \cdots, w_p^T x$ after every iteration.

3. Denoising by Wavelet Thresholding

Removing noise components by thresholding the wavelet coefficients is based on the assumption that in a noisy speech signal, the energy of the speech signal is mainly concentrated in a small number of wavelet dimensions [3]. The energy of these coefficients has larger values compared with other coefficients (especially noise) that have their energy spread over a large number of wavelet coefficients [3]. Thresholding the smaller wavelet coefficients to zero may reduce the noise components from a noisy speech signals [3].

Level dependent wavelet threshold techniques are used widely to suppress the noise from the noisy speech signal and improve the intelligibility of the speech [3]. This method is used in this paper because the forensic audio recording is corrupted with different types of colored noise and these noises have different distributions in different frequency subbands. Thresholding the wavelet coefficients for high frequencies(detail) of the noisy speech signal at each level may reduce the effect of the colored noise at high levels of noise. Level dependent threshold (λ) can be represented by [3]:

$$\lambda = \sigma_j(\sqrt{2\log N_j}) \tag{6}$$

$$\sigma_j = \frac{\mathrm{MAD}(D_j)}{0.6745} \tag{7}$$

where MAD is the median absolute deviation of the detailed coefficients for each level (D_j) and N_j is the length of the noisy speech signal for each level.

The procedure of level dependent wavelet threshold techniques can be illustrated by the following steps.

- Frame the noisy speech signal into several segments by using a Hamming window. The frame duration used in this paper is 25 msec.
- Compute the wavelet coefficients of the noisy speech signal by using discrete wavelet transform (DWT).
- Threshold the detailed coefficients of the noisy speech signal by using a hard or a soft level dependent threshold. Hard (T_{hard}) and soft (T_{soft}) thresholds can be expressed by the following equations:

  $$T_{hard}(D_j) = \begin{cases} D_j, |D_j| > \lambda \\ 0, |D_j| \leqslant \lambda \end{cases} \tag{8}$$

  $$T_{soft}(D_j) = \begin{cases} sign(D_j) * (|D_j| - \lambda), |D_j| > \lambda \\ 0, |D_j| \leqslant \lambda \end{cases} \tag{9}$$

- Reconstruct the enhanced speech signal by applying the inverse discrete wavelet transform to the thresholded wavelet coefficients.

4. Spectral Subtraction

This method is based on subtracting the estimated power spectrum of the noise from the power spectrum of the noisy speech signal, without prior knowledge of the power spectral density of the clean speech and noise signals. Spectral subtraction can be used to suppress background noise by assuming the noise is stationary or changing slowly during the non-speech and speech activity periods [2].

The procedure of spectral subtraction can be summarized by the following steps. Firstly, the noisy speech signal is framed into several overlapping segments by using a Hamming window. The duration of the frame used in this paper is 25 msec and the duration of the overlap between two successive windows is 12.5 msec [10]. Secondly, the noise is estimated by computing the average power spectrum of noise from several silence frames (noise only). Spectral distance voice activty detector is used to determine the noise frames. Then, Fourier transform has been applied to the windowed frames of the noisy speech signal . Spectral subtraction can be computed as [10]:

$$|\hat{S}(k)|^2 = \begin{cases} |X(k)|^2 - \delta|\hat{D}(k)|^2, |X(k)|^2 - \delta|\hat{D}(k)|^2 > \beta|\hat{D}(k)|^2 \\ \beta|\hat{D}(k)|^2, \text{Otherwise} \end{cases} \tag{10}$$

where $X(k)$, $\hat{S}(k)$ and $\hat{D}(k)$ are the magnitude power spectrum of the segment of corrupted speech, estimated speech and estimated noise respectively, δ is the over subtraction factor and it depends on a posteriori segmental SNR, and β is the spectral factor with values between 0 and 1. For a large value of β, the spectral floor is high and the remaining noise is audible, while choosing a small value of β, the noise is significantly reduced, but the remnant noise becomes annoying. Hence,the optimum value of β used in this paper is 0.03 [10]. Finally, the enhanced speech signal $\hat{s}(t)$ can be obtained by applying an inverse Fourier transform to the phase function of discrete Fourier transform of the input speech signal and the estimated spectrum of the speech $|\hat{S}(k)|$.

5. Simulation Results

In this section, we present the simulation results of the independent component analysis, as well as a comparison with spectral subtraction and wavelet denoising techniques for the speech enhancement algorithms. For this paper, 4 levels and Daubechies 8 of the wavelet family were used, respectively. One hundred sentences from forensic voice comparison databases were used for simulation. The forensic voice comparison databases Australian English: 500+ speakers [11] consisted of 532 Australian speakers. Each speaker was recorded in three speaking styles (informal telephone conversation, information exchange task over the telephone and pseudo police style) which are popular speaking styles in forensic applications. The speech was sampled at 44.1 kHz and 16 bit/sample resolution in this database.

Various types of real environmental noise were used in this test from NOISEX-92 [12] and QUT-NOISE databases [13]. The NOISEX-92 database consists of various types of real environmental noise, recorded at 19.98 kHz sample rate with 16 bit resolution [12]. The QUT-NOISE database was created by collecting 10 hours of background noise in 5 common scenarios (cafe, home, street, car and reverberation). Each type of noise was recorded in two locations and the noise signal was sampled at 48 kHz sample rate with 16 bit resolution [13].

In these simulated results, the first and second microphones (x_1 and x_2) have the same distance to the clean speech source from forensic voice comparison database, but the noise(car, street noise from QUT-NOISE database and factory noise from NOISEX-92 database) has different distance to the second microphone resulting in the value of the mixing coefficient of the noise (α). These noises were used in this paper because these types of real environmental noise are more likely to occur in real covert forensic recordings. The mixed speech signal in an ICA algorithm can be represented by :

$$x_1 = s(n) + e(n) \tag{11}$$

$$x_2 = s(n) + \alpha e(n) \tag{12}$$

where $s(n)$ is the original speech signal and $e(n)$ is the noise.

Two down sampling frequencies were used in this paper. Firstly, the car and street noises were down sampled to 44.1 kHz before mixing with clean speech signal. Secondly, the speech signal was also down sampled to 19.98 kHz when factory noise was corrupted with clean speech signal. The down sampled is necessary to match the sampling frequencies of the clean speech and noise signals.

The mixed speech signals are separated using the fast ICA algorithm and the contrast function used in fast ICA has a Gaussian function [8]

There is an arbitrariness in the sign upon inversion. The problem of the sign change of the samples of estimated speech compared with samples of original speech in an ICA algorithm is solved by multiplying all samples of the estimated speech signal by -1 if the maximum correlation coefficient between estimated and original speech has a negative sign.

To evaluate the performance of speech enhancement algorithms in removing the noise from the speech signal, we use the reconstruction SNR or SNR output. It is defined as follows [3]:

$$\mathrm{SNR_o} = \frac{\sum\limits_n s^2(n)}{\sum\limits_n |s(n) - \hat{s}(n)|^2} \tag{13}$$

where $s(n)$ is the original speech signal, and $\hat{s}(n)$ is the estimated original speech signal. The SNR enhancement($\mathrm{SNR_e}$) in (dB) can be defined by:

$$\mathrm{SNR_e} = \mathrm{SNR_o} - \mathrm{SNR_i} \tag{14}$$

where $\mathrm{SNR_i}$ is the input SNR and it can be computed by the ratio of the sum squared of the clean speech to that of the noise from the first microphone (x_1).

To evaluate the effect of the changing mixing coefficient (α) on the performance of ICA to separate the noise from the noisy speech signal, we chose different values of α, ranging from 0.4 to 2.0. Experimental results demonstrated that increasing the value of α decreased the average SNR enhancement when car, street and factory noise were added to 100 sentences from forensic voice comparison database.

Figures (1-3) show comparisons of the average and standard deviation of SNR enhancement for different speech enhancement algorithms when 100 sentences from the forensic voice comparison database were corrupted with street noise, factory noise and car noise. Standard deviations from the Monte Carlo simulation are given on the bars. The value of (α) used in the simulation results of Figures (1-3) was 2 to compare the performance of multiple speech enhancement algorithm (ICA) under worst case conditions with single channel algorithms.

From Figures (1 to 3) we conclude the following:

- Independent component analysis achieves significant improvement in average SNR, compared with spectral subtraction and wavelet threshold techniques, when the speech signals were corrupted with street, car and factory noise for input SNR ranging from -10 to 10 dB.
- Level dependent wavelet denoising techniques achieve higher average SNR enhancement compared with the spectral subtraction algorithm for the same conditions, because real environmental noise are not uniformly distributed over the whole frequencies. Thresholding the detail coefficients in each high frequency sub band by using level dependent wavelet threshold will lead to improved average SNR enhancement at high levels of noise.

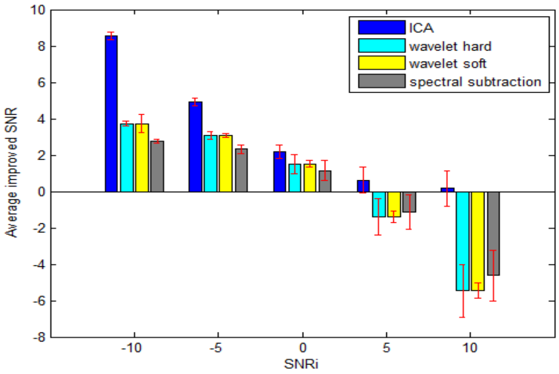


Figure 1: *Comparison of average SNR enhancement when street noise is added to the forensic database*

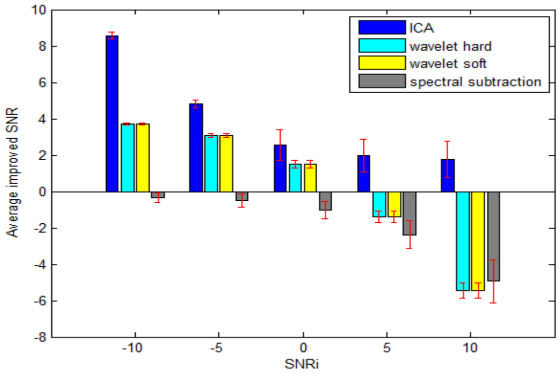


Figure 3: *Comparison of average SNR enhancement when car noise is added to the forensic database*

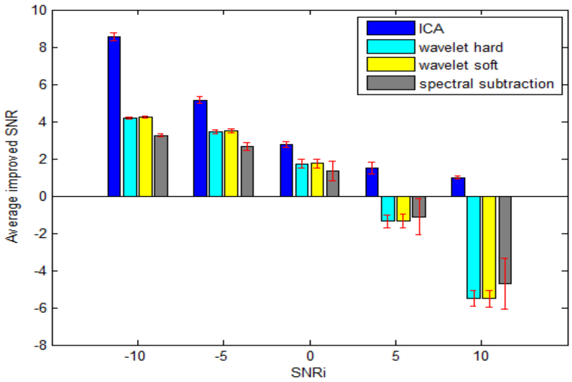


Figure 2: *Comparison of average SNR enhancement when factory noise is added to the forensic database*

- Level dependent threshold and spectral subtraction fails to suppress real environmental noise for input SNR in the range 5 to 10 dB, because power spectral densities of real environmental noise are concentrated at certain frequencies. Using a fixed oversubtracting parameter in spectral subtraction or thresholding all high frequency sub bands of the noisy speech signal at low levels of noise will lead to a distortion of the enhanced speech signal.

6. Conclusions

This paper compares the performance of ICA with specral subtraction and wavelet level dependent threshold techniques to suppress real environmental noise from noisy forensic recordings. Computer simulation results show that ICA achieves higher average SNR improvement than single speech enhancement algorithms for SNR levels in the range -10 dB to 10 dB. Further work is required to investigate the effect of channel delay duration on the performance of the convolutive ICA to suppress the noise from noisy speech signal and compare this result with other single speech enhancement algorithm (Wiener filter) and multiple speech enhancement (beamforming algorithm) for forensic applications.

7. References

[1] Denk, F., da Costa, J. P. C. L. and Silveira, M. A., "Enhanced forensic multiple speaker recognition in the presence of coloured noise", 8th IEEE Int. Conf. Signal Process. Commun. Syst., 2014, pp. 1-7.

[2] Berouti , M., Schwartz, R. and Makhoul, J., "Enhancement of speech corrupted by acoustic noise", IEEE Int. Conf. Acoust., Speech, Signal Process., vol. 4, 1979, pp. 208-211.

[3] Ghanbari , Y. and Karami-Mollaei, M.R., "A new approach for speech enhancement based on the adaptive thresholding of the wavelet packets", Speech Commun., vol. 48, no. 8, pp. 927-940, 2006.

[4] Ruwei , L., Changchun, B., Bingyin, X. and Maoshen, J., "Speech enhancement using the combination of adaptive wavelet threshold and spectral subtraction based on wavelet packet decomposition", 11 th IEEE Int Conf. Signal Process., 2012, pp. 481-484.

[5] Donho, D.L and Johnston, I.M., "Ideal spatial adapation by wavelet shrinkage", Biometrika J., vol. 81, pp. 425-455,1994.

[6] Cao, Y., Sridharan, S. and Moody, M.P., "Post-microphone-array speech enhancement with adaptive filters for forensic application", Int. Symp. Speech, Image Process. Neural Netw., 1994, pp. 253-255.

[7] Zou , X., Jancovic, P., Liu, J. and Kokuer, M., "Speech Signal Enhancement Based on MAP Algorithm in the ICA Space", IEEE Trans. Signal Process., vol. 56, no. 5, pp. 1812-1820, 2008.

[8] Hyvarinen, A. and Oja, E., "Independent component analysis: algorithms and applications", Neural Netw., vol. 13, no. 4, pp. 411-430, 2000.

[9] Singh, L. and Sridharan, S., "Speech enhancement for forensic applications using dynamic time warping and wavelet packet analysis", 10 Annual Conf. IEEE Region Speech Image Technologies Computing Telecommun., vol. 2, 1997, pp. 475-478.

[10] Upadhyay, N. and Karmakar, A., "Speech Enhancement using Spectral Subtraction-type Algorithms: A Comparison and Simulation Study", Procedia Comput. Sci., vol. 54, pp. 574-584, 2015.

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

[12] Varga, A. and Steeneken, H.J., "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 Commun., vol. 12, no. 3, pp. 247-251, 1993.

[13] Dean, D.B., Sridharan, S., Vogt, R.J. and Mason, M.W., "The QUT-NOISE-TIMIT corpus for the evaluation of voice activity detection algorithms", Proc. Interspeech, Makuharia, Japan,2010, pp. 26-30.

A Kalman filtering algorithm with joint metrics-based tuning for single-channel speech enhancement

Aidan E.W. George, Stephen So, Ratna Ghosh, Kuldip K. Paliwal

Signal Processing Laboratory, Griffith School of Engineering,
Griffith University, Brisbane, QLD, Australia, 4111.
{a.george, s.so, k.paliwal}@griffith.edu.au, rg@iee.jusl.ac.in

Abstract

In this paper, we present an iterative Kalman filtering algorithm that exhibits better speech enhancement by jointly utilising robustness and sensitivity metrics. Typically, poor model parameter estimates lead to a biased Kalman filter gain, which results in innovation noise 'leaking' into the output. In the proposed algorithm, the Kalman filter gain is dynamically tuned based on a varying operating point of balanced robustness and sensitivity. Speech enhancement experiments showed the proposed Kalman filtering algorithm to produce higher quality speech than conventional methods using objective and subjective measures.

Index Terms: Speech enhancement, Kalman filtering, noise reduction

1. Introduction

The role of speech enhancement is to reduce the level of undesirable background noise in digitally-recorded speech in order to improve its quality and intelligibility. Several speech enhancement algorithms have been reported in the literature that have had varying degrees of success in enhancing speech but many of them suffer from problems with residual noise, such as the musical noise typically present in Wiener filtering and spectral subtraction algorithms [1]. The Kalman filter was first applied to speech enhancement by Paliwal and Basu [2], and since then has been investigated in the literature both in the time-domain (e.g. [3, 4, 5]) and the modulation-domain [6].

The Kalman filter is an unbiased and linear minimum-mean-squared-error (MMSE) estimator [7] that estimates the clean state vector of speech samples by using a weighted combination of predictions from a speech production model and noise-corrupted speech measurements. The autoregressive (AR) model of speech is commonly used with Kalman filtering to provide the predicted component. While performing remarkably well in the *oracle* case, where AR parameters from the clean speech were available [2], the Kalman filter exhibits poor enhancement performance in practice (non-oracle), where only the noise-corrupted speech is available. This is because the presence of noise leads to bias in the AR parameter estimates. This has a detrimental effect on the enhancement ability in the regions where speech is absent, since the AR estimation bias offsets the *Kalman filter gain*, which regulates how much of the (noisy) innovation signal is used to correct the AR model prediction [5]. This Kalman filter gain offset in the speech-absent regions results in noise from the innovation signal 'leaking' into the output.

In this paper, we propose a new Kalman filtering algorithm that reduces the detrimental effects of poor AR parameter estimates on enhancement performance by jointly utilising two metrics (robustness and sensitivity) [8] that are computed in real-time. The algorithm is iterative in nature. In the initial iteration, the Kalman filter is operated in a constrained mode, where the sensitivity and robustness metrics are balanced. Then, in the subsequent iteration, the AR parameters are estimated from the pre-processed speech and then used in a delayed Kalman filter [2]. Experiments were performed on speech that was corrupted with white Gaussian noise (WGN). The experimental results presented in this paper show that the proposed algorithm exhibited a large improvement in performance over the normal (non-oracle) iterative Kalman filter[1].

2. Kalman filter-based speech enhancement

The additive noise model generally assumed in the problem of speech enhancement is:

$$y(n) = x(n) + v(n) \tag{1}$$

where $x(n)$ is the clean speech, $v(n)$ is a white Gaussian noise (with a variance of σ_v^2), and $y(n)$ is the noise-corrupted speech. The speech and noise signals are assumed to be zero-mean and uncorrelated with each other. In the Kalman filter, a pth order autoregressive (AR) model is used to represent speech production, whose parameters $\{a_k; k = 1, 2, \dots, p\}$ and σ_w^2, represent the AR coefficients and excitation noise variance, respectively [5].

The Kalman filter recursively computes an *a posteriori* state vector estimate $\hat{\boldsymbol{x}}(n|n)$ at time n, when given the noisy speech measurement $y(n)$ and the *a priori* state vector estimate $\hat{\boldsymbol{x}}(n|n-1)$. For a detailed description of each variable, the reader is referred to [5, 7]:

$$\boldsymbol{P}(n|n-1) = \boldsymbol{A}\boldsymbol{P}(n-1|n-1)\boldsymbol{A}^T + \sigma_w^2 \boldsymbol{d}\boldsymbol{d}^T \tag{2}$$

$$\boldsymbol{K}(n) = \boldsymbol{P}(n|n-1)\boldsymbol{c}\left[\sigma_v^2 + \boldsymbol{c}^T\boldsymbol{P}(n|n-1)\boldsymbol{c}\right]^{-1} \tag{3}$$

$$\boldsymbol{P}(n|n) = [\boldsymbol{I} - \boldsymbol{K}(n)\boldsymbol{c}^T]\boldsymbol{P}(n|n-1) \tag{4}$$

$$\hat{\boldsymbol{x}}(n|n-1) = \boldsymbol{A}\hat{\boldsymbol{x}}(n-1|n-1) \tag{5}$$

$$\hat{\boldsymbol{x}}(n|n) = \hat{\boldsymbol{x}}(n|n-1) + \boldsymbol{K}(n)[y(n) - \boldsymbol{c}^T\hat{\boldsymbol{x}}(n|n-1)] \tag{6}$$

In conventional Kalman filtering, the enhanced speech signal at the present time n is given by[2]:

$$\hat{x}(n|n) = \boldsymbol{c}^T\hat{\boldsymbol{x}}(n|n) \tag{7}$$

This is equivalent to taking the first scalar component of the state vector $\hat{\boldsymbol{x}}(n|n)$. Therefore, it is possible to re-write some of the Kalman recursion equations in scalar form.

$$\begin{aligned}\hat{x}(n|n) &= \hat{x}(n|n-1) + K(n)[y(n) - \hat{x}(n|n-1)] &(8)\\ &= [1 - K(n)]\hat{x}(n|n-1) + K(n)y(n) &(9)\end{aligned}$$

where $\hat{x}(n|n-1)$ and $K(n) = \boldsymbol{c}^T\boldsymbol{K}(n)$ are the first scalar components of the *a priori* state vectors and Kalman gain vector, respectively. It can be seen that the scalar Kalman filter gain adjusts the relative proportions of the noisy observation sample

[1] The MATLAB code and sample speech output files are available at http://tiny.cc/speech_enhancement

[2] We used **bold** variables to denote matrix/vector quantities, as opposed to unbolded variables for scalar quantities.

$y(n)$ and *a priori* sample $\hat{x}(n|n-1)$.

We can re-write Eq. (3) in terms of scalar quantities [9]:

$$K(n) = \frac{\alpha^2(n) + \sigma_w^2}{\alpha^2(n) + \sigma_w^2 + \sigma_v^2} \tag{10}$$

where $\alpha^2(n) = \boldsymbol{c}^T \boldsymbol{A}\boldsymbol{P}(n-1|n-1)\boldsymbol{A}^T \boldsymbol{c}$ represents the contribution of the *a posteriori* mean squared error from the previous time step $n-1$, to the total *a priori* mean squared error of the speech model prediction.

Equation (10) provides further insight into the operation of the Kalman filter when used in speech enhancement. When there is no corrupting noise (i.e. $\sigma_v^2 = 0$), the scalar Kalman filter gain becomes unity and therefore, when substituted into Eq. (9), the enhanced speech sample is formed from the observed speech sample $y(n)$ only. On the other hand, when the corrupting noise levels are much higher than the *a priori* mean squared error (i.e. $\sigma_v^2 \gg \alpha^2(n) + \sigma_w^2$), the Kalman filter will favour the predicted speech sample computed from the speech production model.

It is known that noise in the speech will add bias to the AR parameter estimates, which in turn will have a degrading effect on the enhancement performance of the Kalman filter. In this study, we focus our attention on the biased excitation variance estimate, $\tilde{\sigma}_w^2$. For corrupting noise that is white and Gaussian, it can be generally assumed[3] that $\tilde{\sigma}_w^2 \approx \sigma_w^2 + \sigma_v^2$. Therefore, after substituting this into Eq. (10), the scalar Kalman filter gain will also become biased:

$$\tilde{K}(n) = \frac{\alpha^2(n) + \tilde{\sigma}_w^2}{\alpha^2(n) + \tilde{\sigma}_w^2 + \sigma_v^2} \tag{11}$$

$$\approx \frac{\alpha^2(n) + \sigma_w^2 + \sigma_v^2}{\alpha^2(n) + \sigma_w^2 + 2\sigma_v^2} \tag{12}$$

The negative effect of the biased scalar Kalman filter gain manifests itself mostly in the silent pauses between spoken words. Since speech is not present in the silent pauses (i.e. $\sigma_w^2 = \alpha^2(n) = 0$), the biased Kalman filter gain will fluctuate around 0.5. According to Eq. (9), this means that half of the corrupting noise $y(n)$ is passed through to the output.

In this paper, we will consider a new algorithm for reducing the effect of biased estimates on the scalar Kalman filter gain. Previous studies have investigated the use of iterative techniques [3] to reduce estimation bias, where the AR coefficients and excitation variance are re-estimated from filtered speech and then are used in the next iteration. In practice, iterative Kalman filters often suffer from speech distortion and musical noise. However, it has been shown that the initial iteration is important and that better parameter estimates in this iteration will generally result in better performance [5]. In this study, we aim to improve the iterative Kalman filter by quantitatively monitoring the effects of estimation bias using two performance metrics and then tuning the Kalman filter gain in the first iteration.

3. Proposed Kalman filtering algorithm

3.1. Joint robustness and sensitivity metrics

Robustness relates to the ability of the Kalman filter to mitigate uncertainty in its dynamic model parameters. A performance metric for measuring the level of robustness in the Kalman filter was proposed in the instrumentation literature [8]. The robustness metric $J_2(n)$ can be rewritten in terms of scalar quantities as:

$$J_2(n) = \frac{\sigma_w^2}{\alpha^2(n) + \sigma_w^2} \tag{13}$$

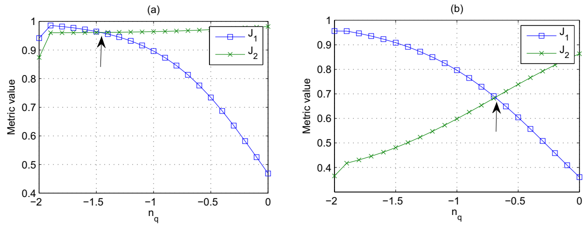


Figure 1: Plot of average Kalman filter sensitivity and robustness metrics (J_1 and J_2) over a frame for varying speech excitation variance $\hat{\sigma}_w^2 = (10^{n_q})\sigma_w^2$ (compromise point where $J_1 = J_2$ shown by the arrow) in a: (a) speech-absent region (only noise); and (b) voiced-speech region.

It can be observed that the speech excitation variance σ_w^2 plays an important role in determining how robust the Kalman filter is. In other words, a large σ_w^2 indicates an unreliable speech model, such as that for voiced speech, which has strong harmonic structure that cannot be predicted by a low-order AR model.

The *sensitivity* metric $J_1(n)$, which quantifies the ability of the Kalman filter to respond to dynamic changes in the input speech (and accordingly, changes in the speech model parameters) in order to mitigate the effects of measurement noise, can be expressed in scalar quantities as:

$$J_1(n) = \frac{\sigma_v^2}{\alpha^2(n) + \sigma_v^2 + \sigma_w^2} \tag{14}$$

The sensitivity metric is dependent on the measurement noise variance σ_v^2. In the case where the measurement signal is too highly corrupted with noise (i.e. $\sigma_v^2 \gg [\sigma_w^2 + \alpha^2(n)]$), the Kalman filter becomes more reliant on its speech model.

The excitation noise variance σ_w^2, which measures the uncertainty of the speech model, appears in both equations for $J_1(n)$ and $J_2(n)$. Changes in this variance term will invariably lead to variations in the metrics [8]. Figure 1 shows a plot of the two metrics (averaged over a frame) as σ_w^2 is varied, i.e. $\hat{\sigma}_w^2 = (10^{n_q})\sigma_w^2$, within regions where there is no speech [Figure 1(a)] and where there is voiced speech [Figure 1(b)]. We can see that the compromise point of *balanced* robustness and sensitivity varies a lot between frames of only noise and frames of voiced speech. Similarly, the values of the metric averages (J_1 and J_2) at this compromise point vary between 0 and 1. The value of $\hat{\sigma}_{wc}^2$ (at this compromise point) can be derived by equating Eqs. (13) and (14):

$$\frac{\sigma_v^2}{\alpha^2(n) + \sigma_v^2 + \hat{\sigma}_{wc}^2} = \frac{\hat{\sigma}_{wc}^2}{\alpha^2(n) + \hat{\sigma}_{wc}^2}$$

Solving for $\hat{\sigma}_{wc}^2$ (and keeping only the positive root), we can compute the excitation variance when the sensitivity and robustness metrics are balanced:

$$\hat{\sigma}_{wc}^2 = \frac{\alpha(n)\sqrt{\alpha^2(n) + 4\sigma_v^2} - \alpha^2(n)}{2} \tag{15}$$

A recent method that used only the robustness metric to tune the Kalman filter was presented in [9]. In this method, the Kalman filter gain was tuned using the value of $J_2(n)$ in order to reduce the effects of bias:

$$K'(n) = [1 - J_2(n)]K(n) \tag{16}$$

One of the problems with this method was that there was over-suppression of the Kalman filter gain even in the speech regions, which introduced speech distortion. The value of $J_2(n)$ (seen in Figure 1 at $n_q = 0$) can be seen to not vary much between noise-only and speech-present frames. In our proposed method,

[3] This assumes that the bias in the AR coefficients a_k is negligible.

Table of Contents for this manuscript

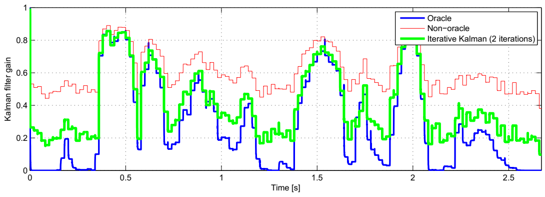


Figure 2: Plot of the scalar Kalman filter gain for oracle and non-oracle modes compared with that of the conventional iterative Kalman filter (with two iterations) [3]. Utterance was sp10 ("The sky that morning was clear and bright blue") corrupted with WGN at 5 dB SNR.

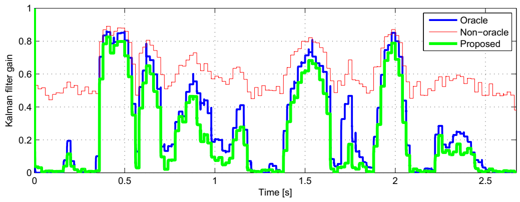


Figure 3: Plot of the scalar Kalman filter gain for oracle and non-oracle modes compared with that of the proposed iterative Kalman filter (with two iterations). Utterance was sp10 ("The sky that morning was clear and bright blue") corrupted with WGN at 5 dB SNR.

instead of using $J_2(n)$ alone in the Kalman filter gain tuning, we utilise $J_{2c}(n)$ ($J_2(n)$ at the compromise point), which combines the effect of the sensitivity metric $J_1(n)$, as was seen in Figure 1. This enables the tuning to better handle both speech-absent and speech-present frames.

We can summarise the proposed iterative Kalman filtering algorithm. For each frame:

Step 1: In the *first iteration*, compute the value of $\hat{\sigma}^2_{wc}$ using Eq. (15);

Step 2: Substitute into Eq. (13) to compute the robustness metric at the compromise point $J_{2c}(n)$;

Step 3: After computing Eq. (13), adjust the Kalman filter gain as in Eq. (16) using $K'(n) = [1 - J_{2c}(n)]K(n)$;

Step 4: In the *second iteration*, estimate the AR parameters from the enhanced speech of the first pass and filter speech using the delayed Kalman filtering algorithm [2] with no Kalman filter gain modification.

3.2. Comparison of the proposed algorithm with the conventional iterative Kalman filter

In this section, we will compare the performance of the proposed iterative Kalman filtering algorithm with the conventional iterative algorithm of Gibson, et al. [3]. For the basis of comparison, both methods use two iterations, with the delayed version of the Kalman filter in the second iteration. Figures 2 and 3 show the scalar Kalman filter gain trajectories for the conventional and proposed iterative Kalman filtering algorithms, respectively. We can see that in the proposed algorithm, the gain is better suppressed in the silence regions, when compared with the conventional one. In the speech regions, the gain in the proposed algorithm is generally similar to what is seen in the oracle case.

4. Experimental setup

The NOIZEUS speech corpus [1] was used in our enhancement experiments, which is composed of 30 phonetically balanced sentences belonging to six speakers. The sampling frequency was 8 kHz. For our objective experiments, we generated a set of stimuli that has been corrupted by WGN at different SNR levels. The segmental signal-to-noise ratio (SNR) and perceptual evaluation of speech quality (PESQ) [1] were used to evaluate the following treatment types.

1. Speech corrupted with white Gaussian noise (**No enhancement, noisy**);
2. Kalman filter (delayed) with AR parameters estimated from clean speech [2] (**Kalman oracle**) ;
3. Kalman filter (iterative) [3] with two iterations (**Kalman iterative**);
4. Proposed Kalman filter with two iterations (**Kalman proposed**); and
5. Minimum mean squared error short-time spectral amplitude estimator [10] (**MMSE-STSA**).

In order to determine the subjective quality of the proposed method in comparison with the other speech enhancement algorithms, a blind AB listening test [4] was performed. Pairs of stimuli were played back to 11 English-speaking listeners, who were then asked to make a subjective preference for each pair. The speech utterance 'She had a smart way of wearing clothes' was corrupted by white Gaussian noise at an SNR of 10 dB. The total number of pair comparisons for all treatment types was 30. This method was preferred over conventional MOS (mean opinion score)-based listening tests, since the scores can have a large variance due to the lack of trained listeners.

5. Results and discussion

Tables 1 and 2 list the average PESQ and segmental SNR of all enhancement methods, respectively. We can see that the Kalman oracle method achieves the highest objective scores since its AR model is estimated from the clean speech. The proposed Kalman filtering algorithm (Kalman proposed) has mostly improved PESQ scores over the Kalman iterative method, especially at the low input SNRs, while it appears to be slightly better than the MMSE-STSA method. However, in terms of segmental SNR (shown in Table 2), the Kalman proposed method can be seen to be significantly outperforming MMSE-STSA and is even competitive with the Kalman oracle method. Recent findings in the speech enhancement literature [11] have found segmental SNR to be more consistent with subjective preference scoring than PESQ. We have similarly found the segmental SNR improvements associated with the proposed Kalman filtering algorithm to be consistent with improved subjective quality in the listening tests that are described later.

In Figure 4, spectrograms are shown of the clean and noisy speech, as well as the output from the four enhancement methods. The utterance was "Clams are small, round, soft, and tasty", where the input SNR is at 10 dB. We can see that the Kalman oracle method in Figure 4(c) exhibited the best enhancement performance, which was consistent in terms of its objective measures (PESQ and segmental SNR). However, when using the AR parameter estimates from the noisy speech, residual noise started appearing in the Kalman iterative method, as shown in Figure 4(d). The proposed method [Figure 4(e)] exhibited a comparable level of residual noise to the oracle method. The MMSE-STSA method [Figure 4(f)] appeared to suffer from a metallic residual noise in all frequency bands, hence it exhibited a lower PESQ and segmental SNR.

Figure 5 shows the mean preference scores from the subjective listening tests as well as error bars that indicate 95% confidence intervals. It can be seen that apart from the clean speech, the Kalman oracle method was the most preferred method by the listeners. The proposed method was the next preferred followed by the MMSE-STSA and Kalman iterative method. As mentioned previously, these subjective preference scores are correlated with the segmental SNR results of Table 2.

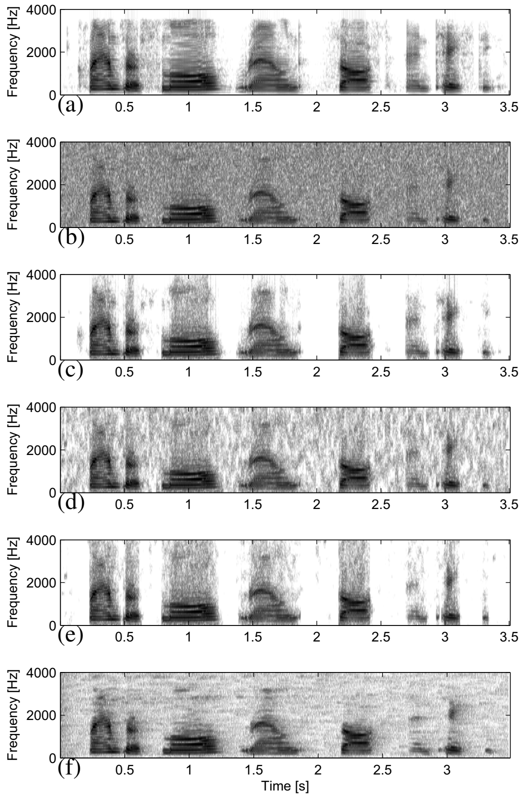


Figure 4: Spectrograms of digital speech ("Clams are small, round, soft, and tasty"): (a) with no noise; (b) corrupted by WGN at 10 dB SNR; (c) enhanced by Kalman oracle; (d) enhanced by Kalman iterative; (e) enhanced by Kalman proposed; and (f) enhanced by MMSE-STSA.

6. Conclusion

In this paper, we have presented an iterative Kalman filtering algorithm that utilises robustness and sensitivity metrics jointly to dynamically tune the Kalman filter gain to overcome poor AR parameter estimates. Both of these metrics are computed in real-time and incorporated into an iterative Kalman filtering framework. Experimental results (PESQ and segmental SNR) showed the proposed method to be competitive with the oracle-case Kalman filter and was better than the MMSE-STSA algorithm. Subjective blind listening tests also corroborated the objective findings, where the listeners preferred the enhanced speech from the proposed method over those produced by the MMSE-STSA algorithm and conventional iterative Kalman filter.

Table 1: Average PESQ results over 30 sentences from the NOIZEUS database, which compare the different speech enhancement methods with the proposed method for speech corrupted by white Gaussian noise.

Method	Input SNR (dB)			
	0	5	10	15
No enhancement	1.57	1.83	2.13	2.47
Kalman oracle	2.50	2.79	3.08	3.38
Kalman iterative	1.92	2.29	2.63	2.98
Kalman proposed	**2.08**	**2.39**	**2.67**	**2.97**
MMSE-STSA	1.96	2.33	2.64	2.94

Table 2: Average segmental SNR (in dB) results over 30 sentences from the NOIZEUS database, which compare the different speech enhancement methods with the proposed method for speech corrupted by white Gaussian noise.

Method	Input SNR (dB)			
	0	5	10	15
No enhancement	−8.31	−3.31	1.69	6.69
Kalman oracle	4.61	6.81	9.41	12.39
Kalman iterative	−0.48	3.48	7.32	11.14
Kalman proposed	**3.32**	**5.85**	**8.68**	**11.78**
MMSE-STSA	−0.32	3.15	6.40	9.40

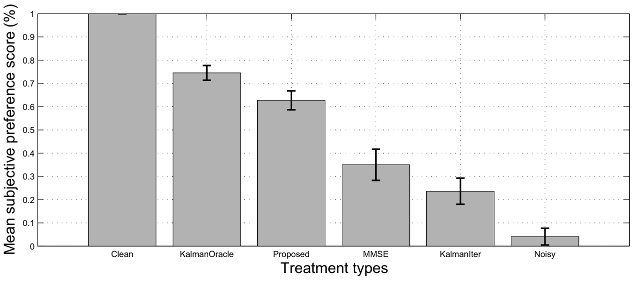


Figure 5: Mean subjective preference scores with 95% confidence intervals for all treatment types.

7. References

[1] P. Loizou, *Speech Enhancement: Theory and Practice*, 1st ed. CRC Press LLC, 2007.

[2] K. K. Paliwal and A. Basu, "A speech enhancement method based on Kalman filtering," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 12, Apr. 1987, pp. 177–180.

[3] J. D. Gibson, B. Koo, and S. D. Gray, "Filtering of colored noise for speech enhancement and coding," *IEEE Trans. Signal Process.*, vol. 39, no. 8, pp. 1732–1742, Aug. 1991.

[4] P. Sorqvist, P. Handel, and B. Ottersten, "Kalman filtering for low distortion speech enhancement in mobile communication," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 2, 1997, pp. 1219–1222.

[5] S. So and K. K. Paliwal, "Suppressing the influence of additive noise on the Kalman filter gain for low residual noise speech enhancement," *Speech Commun.*, vol. 53, no. 3, pp. 355–378, Mar. 2011.

[6] ——, "Modulation-domain Kalman filtering for single-channel speech enhancement," *Speech Commun.*, vol. 53, no. 6, pp. 818–829, Jul. 2011.

[7] M. H. Hayes, *Statistical Digital Signal Processing and Modeling*. New Jersey: John Wiley, 1996.

[8] M. Saha, R. Ghosh, and B. Goswami, "Robustness and sensitivity metrics for tuning the extended Kalman filter," *IEEE Trans. Instrum. Meas.*, vol. 63, no. 4, pp. 964–971, Apr. 2014.

[9] S. So, A. E. W. George, R. Ghosh, and K. 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. 1, pp. 263–268, Aug. 2016.

[10] 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. 32, pp. 1109–1121, Dec. 1984.

[11] B. Schwerin and K. K. Paliwal, "Using STFT real and imaginary parts of modulation signals for MMSE-based speech enhancement," *Speech Commun.*, vol. 58, pp. 49–68, Mar. 2014.

Table of Contents for this manuscript

A comparison of estimation methods in the discrete cosine transform modulation domain for speech enhancement

Aidan E.W. George, Christine Pickersgill, Belinda Schwerin, Stephen So

Griffith School of Engineering, Gold Coast campus,
Griffith University, QLD, 4222 Australia.

{aidan.george, christine.pickersgill}@griffithuni.edu.au,
{b.schwerin, s.so}@griffith.edu.au

Abstract

In this paper, we present a new speech enhancement method that processes noise-corrupted speech in the discrete cosine transform (DCT) modulation domain. In contrast to the Fourier transform, the DCT produces a real-valued signal. Therefore, modulation-based processing in the DCT domain may allow both acoustic Fourier magnitude and phase information to be jointly estimated. Based on segmental SNR and the results of blind subjective listening tests on various coloured noises, the application of the subspace method in the DCT modulation domain processing was found to outperform all other methods evaluated, including the LogMMSE method.

Index Terms: speech enhancement, discrete cosine transform, modulation domain, subspace method

1. Introduction

Environmental noise is an ever-present problem in many speech processing applications, such as speech coding and automatic speech recognition (or ASR), since its presence in the captured speech signal adversely impacts on performance. For instance, in speech coding, the quality and intelligibility of the decoded speech degrades as the signal-to-noise ratio (or SNR) of the input signal decreases. In ASR tasks, low input SNR leads to a deterioration of the recognition accuracy.

One technique for addressing the problem of noise is to use speech enhancement. The role of speech enhancement is to reduce the level of undesirable noise in a speech signal in order to improve the quality and intelligibility. Several speech enhancement algorithms have been reported in the literature, each of which process speech in different domains. Early enhancement methods, such as spectral subtraction [1], MMSE-STSA [2], and LogMMSE [3], estimated the clean speech in the acoustic magnitude domain. In subspace enhancement methods, a linear estimator is applied in the signal and noise subspace domains. Recently, more successful speech enhancement methods have been reported that apply spectral subtraction [4], MMSE [5] and Kalman filters [6] in the modulation magnitude domain.

Typically, in modulation domain methods, an analysis-modification-synthesis (or AMS) procedure is first performed and estimation is applied on the temporal trajectories of Fourier transform magnitudes at each acoustic frequency [4]. The enhanced acoustic Fourier magnitude information is then combined with the noisy acoustic Fourier phase and synthesised using a windowed overlap-adding procedure to obtain the enhanced speech frame. A common characteristic of the acoustic and modulation domain methods is the use of noisy and unprocessed Fourier phase information in the synthesis stage. A recent study [7] showed that at low instantaneous spectral SNR that are less than 7 dB, noise in the phase information has a notable effect on the quality and intelligiblity of the enhanced speech output. Therefore, it is advantageous to process both Fourier magnitude and phase information.

One method of incorporating phase information into the enhancement procedure is to process real and imaginary parts (RI) of the complex Fourier transform [8, 9], which is also referred to as RI modulation. This is based on the concept that the real and imaginary parts contribute to both the magnitude and phase components. In these methods, an estimator is *applied independently* on the real and imaginary modulation signals over time. However, one may speculate that *joint processing* of RI modulation signals is advantageous in better estimating the clean phase information.

In this paper, we report on a new speech enhancement approach that estimates modulation magnitude and phase information in the discrete cosine transform (DCT) domain. It is well known that the DCT is a unitary transform that outputs a real-valued signal, in contrast to the discrete Fourier transform, which outputs a complex-valued signal. This suggests that the Fourier magnitude and phase information is embedded within the single real-valued signal. Therefore, a speech enhancement approach that is based on processing in the DCT modulation domain can be viewed as a *joint estimator* of magnitude and phase information. In order to determine an effective estimation method in the DCT modulation domain, our study evaluated three methods: spectral subtraction [1], LogMMSE [3] and the subspace approach [10]. Speech enhancement experiments on the NOIZEUS speech corpus [11] were performed to evaluate the different methods in the DCT modulation domain as well as compare them against the conventional subspace and acoustic magnitude domain methods, such as spectral subtraction and LogMMSE[1]. Objective scores and blind subjective listening tests suggested that the subspace approach was the more effective method for estimation in the DCT modulation domain. Furthermore, they indicated that the proposed DCT-SSp enhancement method produced better quality enhanced speech than the spectral subtraction, LogMMSE and the conventional subspace method.

2. Method

In this work we propose a modulation-domain speech enhancement method which makes use of the discrete cosine transform (DCT) in place of the short-time Fourier Transform (STFT). The DCT has the advantage in that it provides a frequency representation of a signal where the magnitude and phase information are jointly represented by a single real-valued spectrum. The DCT has been particularly popular in speech and image coding due to its ability to concentrate signal energy in a few coefficients, but has also been applied to speech enhancement. For example, in [12] the MMSE method was used to estimate the DCT coefficients on a frame-by-frame basis. In contrast, we have investigated the use of DCT in a modulation-domain based framework where DCT coefficient time trajectories are processed in order to suppress noise. Further, we demonstrate that the subspace enhancement approach is well suited to enhancing speech in this domain.

[1] The MATLAB code and sample speech output files are available at http://tiny.cc/speech_enhancement

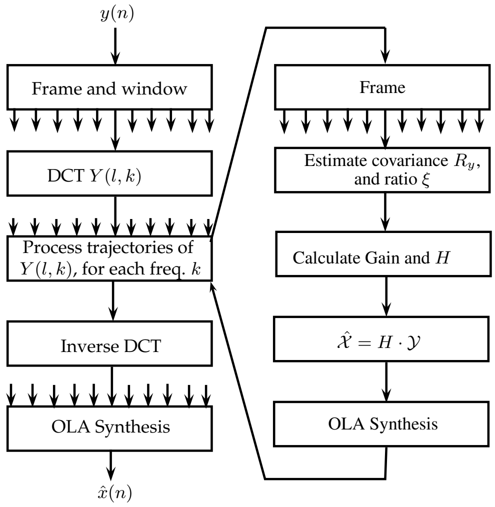

$y(n)$
Frame and window
DCT $Y(l,k)$
Process trajectories of $Y(l,k)$, for each freq. k
Inverse DCT
OLA Synthesis
$\hat{x}(n)$
Frame
Estimate covariance R_y, and ratio ξ
Calculate Gain and H
$\hat{\mathcal{X}} = H \cdot \mathcal{Y}$
OLA Synthesis

Figure 1: *DCT-SSp method for enhancing speech using a DCT-based modulation framework (left) and suppression of noise in the DCT spectral time trajectories using a subspace estimation method (right).*

2.1. DCT-based modulation domain framework

Consider a speech signal $x(n)$ corrupted by additive noise $d(n)$ to give the noisy signal $y(n) = x(n)+d(n)$. Assuming speech to be quasi-stationary, the noisy speech signal can be analysed framewise by applying the DCT to each windowed frame of the signal, to give the noisy DCT spectrum

$$Y(l,k) = u(k)\sum_{n=0}^{N-1} y(n+lZ)v_a(n)\cos\left[\frac{\pi(2n-1)k}{2N}\right]$$
$$k = 0, 1, ..., N-1, \quad (1)$$

where

$$u(k) = \begin{cases} \frac{1}{\sqrt{N}}, & k = 0, \\ \sqrt{\frac{2}{N}}, & 1 \le k \le N-1. \end{cases} \quad (2)$$

Here l refers to the frame index, k is the frequency index, Z is the frame shift, N is the frame length, and $v_a(n)$ is the analysis window.

The time trajectories for each frequency component of the DCT spectra $Y(l,k)$ form a modulation signal which can then be processed framewise to suppress noise. The enhanced speech signal $\hat{x}(n)$ can then be reconstructed from the modified spectra $\hat{X}(l,k)$ using an inverse DCT, followed by overlap-add (OLA) synthesis. That is,

$$\hat{x}(n) = \sum_{l} v_s(n-lZ)\sum_{k=0}^{N-1} u(k)\hat{X}(l,k)\,\cos\left[\frac{\pi(2n-1)k}{2N}\right],$$
$$k = 0, 1, ..., N-1, \quad (3)$$

where $v_s(n)$ is the synthesis window.

2.2. Subspace enhancement of the modulation signal

To enhance speech within the DCT-based modulation framework described in Section 2.1, each modulation signal is processed within a separate framework where the signal is framed and the generalized subspace method for enhancing speech corrupted by coloured noise [10] is used, as represented in Figure 1.[2] Processing of the modulation signal begins with overlapped framing, then the following procedure is used to modify each frame of the signal.

The covariance matrix R_y of each noisy frame is estimated using the multi-window method [13] and sine tapers [14]. For this approach, the number of tapers used has a considerable effect on the resulting quality of enhanced stimuli. Using only a small number of tapers results in increased variance in the estimate, while a higher number of tapers results in reduced variance but reduced resolution (or larger bias). Thus, the number of tapers used in estimating the covariance matrix provides a trade-off between different distortion types.

The ratio of the clean to noise covariance ξ is then estimated from the noisy and noise covariance matrices as

$$\xi = R_n^{-1}R_x = R_n^{-1}R_y - I, \quad (4)$$

where the noise covariance matrix R_n is initially estimated from the leading noise-only region of the utterance (using the multi-window method), and updated in non-speech frames (as determined by a simple SNR-based VAD calculated from the first coefficient of the noisy and noise covariance matrices).

Eigen-decomposition is then used to find the eigenvalues Λ and eigenvectors V of ξ,

$$\xi V = V\Lambda. \quad (5)$$

The dimension of the speech signal subspace M is then given by the number of non-zero positive eigenvalues of ξ.

The gain matrix can then be calculated as

$$g_{kk} = \frac{\lambda(k)}{\lambda(k)+\mu}, \qquad k = 1, 2, ..., M$$
$$G = \text{diag}\{g_{11}, g_{22}, ..., g_{MM}\} \quad (6)$$

where $\lambda(k)$ are the positive non-zero eigenvalues of ξ (in decreasing order). In Eq. (6), μ provides a trade-off between speech distortions and residual noise, and is selected based on SNR as

$$\mu = \begin{cases} \mu_o - \frac{\text{SNR}_{dB}}{slope}, & -5 < \text{SNR}_{dB} < 20 \\ 1, & \text{SNR}_{dB} \ge 20 \\ 5, & \text{SNR}_{dB} \le -5. \end{cases} \quad (7)$$

where SNR_{dB} is estimated from the eigenvalues corresponding to the speech signal subspace (and therefore estimating the eigenvalues of R_x), and μ_o and *slope* are empirically determined values. Here, $\mu_o = 4.2$ and $slope = 0.16$, as suggested in [10] were used.

The estimate of the enhanced frame is then given by

$$\hat{\mathcal{X}}(\ell, m, k) = H \cdot \mathcal{Y}(\ell, m, k)v_m(k), \quad (8)$$

where

$$H = V^{-T}\begin{bmatrix} G & 0 \\ 0 & 0 \end{bmatrix}V^T, \quad (9)$$

and where ℓ is the modulation signal frame index, m is the frame value index, and $v_m(k)$ is the window function. Finally, OLA synthesis is used to reconstruct the modified modulation signal $\hat{X}(l,k)$.

3. Experiments

A number of experiments were conducted to evaluate the quality of stimuli processed using the proposed DCT-SSp method. Both blind subjective tests and objective evaluations were performed. Degraded stimuli used in experiments were stimuli from the Noizeus speech corpus [11] corrupted by F16, Babble and Car noises, at input noise levels ranging from 0 to 15 dB. For subjective evaluations, blind AB listening tests were conducted, in which 16 English-speaking listeners compared the quality of stimuli corrupted by F16 noise at 5 dB and processed using different treatment types, and selected the stimuli they preferred. These listening tests made use of two sentences, one from a male and one from a female speaker, and involved listening to 84 stimuli pairs per experiment. Objective experiments compared stimuli processed using each treatment type to the original clean stimuli, for a range of input SNR values. Mean

[2] For a more detailed account of the generalized subspace approach for speech corrupted by colour noise, the reader is referred to [10].

Table 1: *Parameters used in implementing the proposed DCT-SSp enhancement method.*

Parameter	DCT	Subspace
Frame duration	20 ms	16 ms
Sub-frame duration	-	2 ms
Frame update	0.625 ms	8 ms
Analysis window	Hamming	-
Synthesis window	modified Hann	Hamming

Segmental SNRs for each treatment type and input SNR were then calculated across the 30 utterances of the Noizeus corpus.

An important parameter in the DCT-SSp method is the number of tapers used in the estimation of the covariance matrix, since the variance-bias tradeoff has a significant effect on the quality. A small number of tapers leads to good speech quality but a high level of residual noise, while a large number of tapers results in a more distorted speech along with reduced levels of residual noise. Blind listening tests performed by a subset of listeners found that 32 and 128 tapers were the most preferred, with some listeners preferring lower residual noise levels and others preferring less speech distortion. We have included both in the formal listening tests, where stimuli generated using 32 tapers are denoted DCT-SSp32 and those generated using 128 tapers are denoted DCT-SSp128. Other parameters used in the construction of DCT-SSp stimuli are shown in Table 1.

In the first round of listening tests, the suitability of the subspace method in the DCT modulation domain was examined. Treatment types included in the subjective tests were the proposed subspace method (DCT-SSp), spectral subtraction (DCT-SpSub), and LogMMSE (DCT-LogMMSE). The DCT-SpSub and DCT-LogMMSE methods utilised the DCT framework described in Section 2.1. Each trajectory was then processed using either the spectral subtraction method of Boll [1] or the Log-MMSE magnitude estimator of Ephraim and Malah [3].

In the second set of listening tests, the proposed DCT-SubSpace method was compared with popular acoustic domain methods including spectral subtraction, LogMMSE, and subspace. Acoustic domain enhancement methods were implemented using publicly available reference implementations [11].

4. Results and discussion

Mean subjective preference scores for the first listening test for utterances sp10 (male speaker) and sp15 (female speaker) are shown in Figure 2. Stimuli included in subjective evaluations were clean, noisy, DCT modulation domain spectral subtraction (DCT-SpSub), LogMMSE (DCT-LogMMSE), and Subspace using 32 tapers (DCT-SSp32) and 128 tapers (DCT-SSp128). Scores shown in Figure 2 indicate that both DCT Subspace treatment types were preferred over DCT LogMMSE and DCT spectral subtraction methods.

Separate one-way ANOVA tests were performed on the male and female spoken stimuli sets to determine the statistical significance of the first round of subjective results. The null hypothesis ("all means are equal") was found to have been rejected (for Sp10, F(6, 105) = 130.26, $p < 0.05$; and for Sp15, F(6, 105) = 110.76, $p < 0.05$). Post hoc analysis using the Tukey's Honestly Significant Difference (HSD) tests performed on each stimuli set found most treatment types to be significantly different, with the exception of the comparison between DCT-SSp32 and DCT-SSp128.

A second listening test was performed using clean, noisy, and stimuli processed with acoustic spectral subtraction (SpecSub), acoustic Log-MMSE (LogMMSE), acoustic Subspace using 16 tapers(Ac-SSp), DCT-SSp32, and DCT-SSp128. Mean subjective preference scores given in Figure 3 show that the two DCT Subspace methods are preferred among listeners. Though LogMMSE also shows a high percentage of preference, the post hoc HSD analysis found that it was significantly different to both the DCT-SSp treatment types. Identical ANOVA and post hoc HSD tests performed on both the male and female spoken stimuli sets. The null hypothesis was found to have been rejected (for Sp10, F(5, 96) = 68.0, $p < 0.05$; and for Sp15, F(5, 96) = 55.67, $p < 0.05$) and post hoc analysis identified significant differences between all treatment types, though not between DCT-SSp32 and DCT-SSp128.

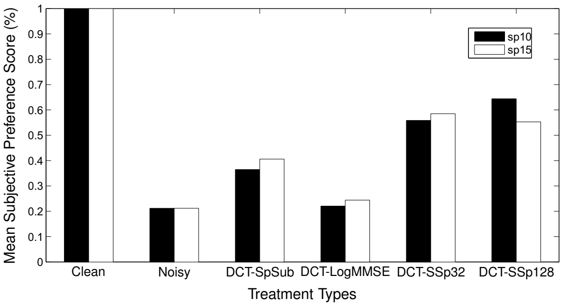


Figure 2: *Average subjective test results featuring clean, noisy (5 dB F16 noise), and stimuli processed with DCT-SpSub, DCT-LogMMSE, DCT-SSp32, and DCT-SSp128.*

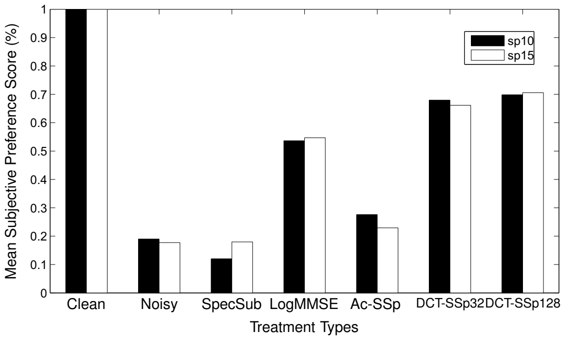


Figure 3: *Average subjective test results featuring clean, noisy (5 dB F16 noise), and stimuli processed with SpecSub, LogMMSE, Ac-SSp, DCT-SSp32, and DCT-SSp128.*

Spectrograms (for utterance sp10) of stimuli used in all subjective quality tests are shown in Figure 4. Clean and noisy (5 dB added F16 noise) are shown in Figures 4(a) and (b) respectively. We can see that the acoustic domain methods (SpecSub and LogMMSE) in Figures 4(c) and (d) suffer from medium levels of residual noise that are characteristic of the respective methods. The conventional Ac-SSp method (Figure 4(e)) exhibits better noise suppression, though the residual noise is rather non-stationary and the speech has suffered from some spectral smoothing, giving it a 'breathy' nature. For the DCT modulation methods, it can be seen that the proposed DCT-SSp32 and DCT-SSp128 methods (Figures 4(h) and (i)) exhibit much better noise suppression than DCT-LogMMSE and DCT-SpSub. The speech is clearer and less bottled for DCT-SSp32, but also contains a residual tone which is not audible in DCT-SSp128. While spectrograms for utterance sp10 are shown here, observations of spectrograms for other utterances from the Noizeus corpus corrupted with 5 dB F16 noise also show similar properties.

Objective quality evaluations were performed for all treatment types at various input SNR levels and for different types of coloured noise. Segmental SNR scores for three noise types (F16, Babble, and Car) are given in Table 2. In most cases the two proposed methods have scored higher than the other treat-

Table 2: *Segmental SNR scores for all treatment types at various input SNR levels and for three types of coloured noise.*

Method	F16				BABBLE				CAR			
	0dB	5dB	10dB	15dB	0dB	5dB	10dB	15dB	0dB	5dB	10dB	15dB
Noisy	-4.329	-1.445	1.763	5.245	-4.098	-1.153	1.972	5.563	-4.253	-1.423	1.807	5.355
SpecSub	-0.149	2.801	6.000	9.307	-0.739	2.132	5.109	8.442	0.178	3.004	6.010	9.464
LogMMSE	0.410	3.051	5.634	8.153	-1.234	1.500	4.200	6.917	0.474	3.001	5.580	8.231
Ac-SSp	0.774	3.237	5.495	7.675	-0.363	2.155	4.408	6.675	1.044	3.302	5.372	7.666
DCT-SpSub	-1.932	-0.870	-0.019	0.573	-3.137	-1.438	-0.695	0.261	-1.711	-0.673	-0.014	0.707
DCT-LogMMSE	-5.199	-5.069	-4.956	-4.828	-5.455	-5.186	-5.080	-4.928	-5.152	-5.028	-4.953	-4.813
DCT-SSp32	1.723	4.283	6.794	9.109	-0.174	2.393	5.312	7.814	1.603	4.066	6.497	8.987
DCT-SSp128	2.428	4.692	7.053	9.214	0.302	2.713	5.584	8.011	2.293	4.437	6.667	9.040

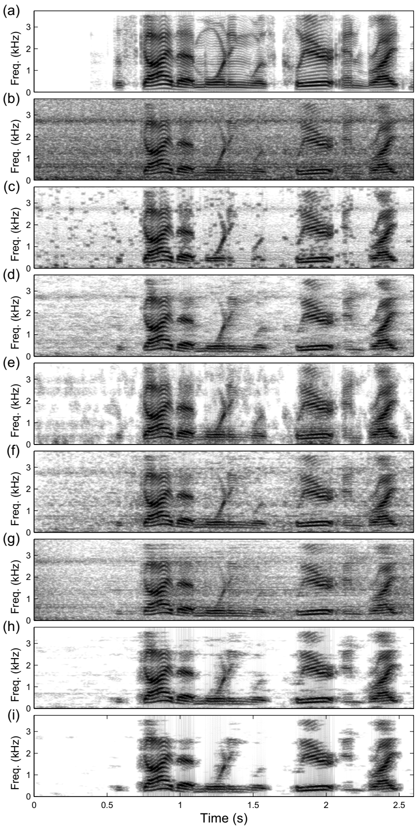


Figure 4: *Spectrograms of sp10 utterance, "The sky that morning was clear and bright blue", by a male speaker from the Noizeus speech corpus: (a) clean speech; (b) speech degraded by 5 dB F16 noise; (c) Spec-Sub; (d) LogMMSE; (e) Ac-SSp; (f) DCT-SpSub; (g) DCT-LogMMSE; (h) DCT-SSp32; (i) DCT-SSp128.*

ment types, often by a considerably large margin. Scores for F16 noise accurately reflect the subjective test results shown in Figure 2 and Figure 3.

5. Conclusion

In this paper, we have investigated a new speech enhancement method that processes noise-corrupted speech in the discrete cosine transform modulation domain. Three enhancement methods (spectral subtraction, LogMMSE, and subspace) that processed DCT spectral coefficients temporally were evaluated using blind subjective listening tests and objective quality measures. It was found that the best enhancement performance was achieved when using the subspace method in the DCT modulation domain.

6. References

[1] 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.

[2] 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.

[3] ——, "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.

[4] K. Paliwal, B. Schwerin, and K. Wójcicki, "Modulation domain spectral subtraction for speech enhancement," in *Proc. ISCA Conf. Int. Speech Commun. Assoc. (INTERSPEECH)*, Brighton, U.K., Sep 2009, pp. 1327–1330.

[5] ——, "Speech enhancement using minimum mean-square error short-time spectral modulation magnitude estimator," *Speech Communication*, vol. 54, no. 2, pp. 282–305, Feb 2012.

[6] 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.

[7] R. Chappel, B. Schwerin, and K. Paliwal, "Phase distortion resulting in a just noticeable difference in the perceived quality of speech," *Speech Communication*, vol. 81, pp. 138–147, July 2016.

[8] Y. Zhang and Y. Zhao, "Spectral subtraction on real and imaginary modulation spectra," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Process. (ICASSP)*, May 2011, pp. 4744–4747.

[9] B. Schwerin and K. Paliwal, "Speech enhancement using STFT real and imaginary parts of modulation signals," in *Proc. of SST*, Sydney, Australia, Dec 2012, pp. 25–28.

[10] Y. Hu and P. Loizou, "A generalized subspace approach for enhancing speech corrupted by colored noise," *IEEE Trans. Speech Audio Process.*, vol. 11, no. 4, pp. 334–341, July 2003.

[11] P. Loizou, *Speech Enhancement: Theory and Practice*. Boca Raton, FL: Taylor and Francis, 2007.

[12] I. Soon, S. Koh, and C. Yeo, "Noisy speech enhancement using discrete cosine transform," *Speech Communication*, vol. 24, pp. 249–257, 1998.

[13] L. McWorter and L. Scharf, "Multiwindow estimators of correlation," *IEEE Trans. Signal Process.*, vol. 46, no. 2, pp. 440–448, Feb 1998.

[14] K. Riedel and A. Sidorenko, "Minimum bias multiple taper spectral estimation," *IEEE Trans. Signal Process.*, vol. 43, no. 1, pp. 188–195, Jan 1995.

Searching for importance: focus facilitates memory for words in English

Heather Kember[1,2] *, Jiyoun Choi*[2,3]*, Jenny Yu*[1,2]

[1]The MARCS Institute, Western Sydney University, Australia
[2]ARC Centre of Excellence for the Dynamics of Language, Australia
[3]Hanyang Phonetics and Psycholinguistics Lab, Hanyang University, Korea

h.kember@westernsydney.edu.au, jiychoi@hanyang.ac.kr, jenny.yu@westernsydney.edu.au

Abstract

Focus, both prosodic and syntactic, conveys processing advantages in English. We compared recognition memory for words that received syntactic, prosodic, syntactic + prosodic, or no focus. English speakers listened to blocks of sentences, then were presented with words and asked if they remembered them. Words with focus were recognised faster than words without focus, and there was an additive effect, such that words with both focus types were responded to even faster than words that had only prosodic, or syntactic focus. These findings suggest a common mechanism for drawing attention towards focused constituents within an utterance.

Index Terms: prosody; focus; memory; speech processing

1. Introduction

Listeners will focus attention on crucial components within an utterance that constitutes newly introduced information [1]. This is considered to be a language universal. In English, marking of focus can be achieved through both syntactic and prosodic means: prosodic focus is marked with a pitch accent [2], and syntactic focus can be achieved via cleft structures such as It-cleft (e.g. it was the…), or There-insertion (e.g. There was this…). Both of these focus-producing devices have been shown to provide processing advantages. For example, words marked by syntactic focus are remembered better compared to words without syntactic focus [3, 4]. As for prosodic focus, sounds within salient words are recognised more rapidly in comparison to sounds within non-salient words [5]. This processing benefit conveyed by prosody is not solely due to the increased acoustic saliency – when the focused words themselves are removed and replaced with neutrally produced words, participants are still faster at detecting phonemes at the prosodically salient position [6, 7]. This suggests that the listener entrains to the utterance prosody, predicting where the focused word will be located, and directs attention to that position.

While these processing benefits for focused constituents within utterances have been documented in English and other languages [8, 9], it is an open question what the mechanism is that is driving this facilitatory processing of focused information, and indeed whether the mechanisms are the same for the different types of focus. [10] suggests a central mechanism, which is a search for the most semantically central part of a speaker's message, which can be provided by any of the focus types. Their paper used semantic focus only however, and not other focus types. Another study [4] presented sentences visually, and tested recognition memory for words that were either identical, phonologically related, or semantically related to target words within the sentence. Overall, words that had syntactic focus were consistently remembered better than words without syntactic focus. Syntactic focus facilitated remembering the identical targets and phonological targets (only with a memory delay) but memory for semantically related targets was not improved. The authors suggested that the memory trace created by syntactic focus is specifically for the identity of the word itself and phonological information, rather than the general meaning of the word. Taken together, these previous studies suggest a central mechanism for seeking the semantically important parts of an utterance, and facilitating processing of the specific word identity.

It is also unclear how syntactic and prosodic focus interacts with each other, as all studies mentioned up to this point, only assessed a single type of focus in an utterance. In [6], the authors compared the effect of semantic and prosodic focus in a phoneme monitoring task. They found that both focus types facilitated phoneme detection over non-focused words, but that there was also an interaction between the two, arguing for a common cause. Given syntactic focus also serves this same purpose as semantic focus (i.e. to alert the listener to important information), we suggest that we will also see this additive effect of focus type in our study.

In this paper we present a pair of experiments investigating the role that syntactic and prosodic focus plays in sentence processing in English. In Experiment 1, we tested recognition memory for words that had been focused (either syntactically, prosodically, or both) in comparison to unfocused words. Experiment 2 replicated Experiment 1, but also manipulated the target words presented, such that targets were either identical or related semantically or phonologically.

2. Experiment 1

Both experiments tested recognition memory for words that had been presented in one of four conditions: no focus, prosodic focus, syntactic focus, or prosodic + syntactic focus. We recorded recognition accuracy and reaction time.

2.1. Method

2.1.1. Participants

32 native speakers of Australian English were recruited for participation (M_{age} = 29.63, SD_{age} = 13.32). Some participants learned other languages (no simultaneous bilinguals). Participants had no self-reported speech, reading, or hearing issues.

2.1.2. Stimuli

140 experimental sentences were constructed. Each sentence contained two possible high frequency target words, one early, and the other later in the sentence. One target had syntactic

focus and the other did not have syntactic focus. No target words were repeated across sentences. We used either the "It was the…" or "There was this…" syntactic cleft structures meaning all syntactically focused words were the first target.

A female, native speaker of Australian English recorded two versions of each sentence. In version one, prosodic focus was placed on the first target, and in version two it was placed on the second target. This generated four different experimental conditions (two per sentence). For example:

1) There was this **medal** (PS) displayed on the stand (NF) in the living room
2) There was this medal (ST) displayed on the **stand** (PR) in the living room

(NF = no focus, PR = prosodic focus, ST = syntactic focus, PS = prosodic + syntactic focus. Prosodic focus indicated by bold type, target word indicated by underline). To encourage the speaker to place focus on the correct target, two questions were written for each sentence and read to the speaker during recording. Participants only heard one version of each sentence to ensure any effect was not simply due to repeated presentations of the same words. To control for the confound of syntactically focused words appearing as target 1, we also created a set of 20 control sentences that also had early and late targets but without syntactic or prosodic focus.

2.1.3. *Procedure*

Sentences were pseudo-randomised so that all sentence types were evenly dispersed across the experiment. Sentences were then split into blocks of 10, with each block containing seven experimental sentences, two filler sentences, and one control sentence. Four different iterations of the experiment were created, varying the order of presentation of stimuli, as well as the sentence versions and targets. For example, in experiment version one the target word "stand" was tested in the no focus condition and in version two, it was tested in the prosodic focus condition. In versions three and four, the target word "medal" was tested in syntactic, and prosodic + syntactic conditions respectively. Participants were randomly assigned to one of the four experimental versions, with numbers equal in each of the groups.

Participants were tested individually in a sound attenuated room in a single session lasting approximately 45 minutes. E-Prime was used to present the stimuli and record participant responses. Noise-attenuating Sennheiser HD 280 Pro headphones were fitted and adjusted to achieve comfortably audible volume levels for each participant individually.

The experiment alternated between blocks of sentences and recognition memory tests. To encourage attentiveness, participants pressed the spacebar to play each sentence. After the 10 sentences were played, instructions appeared on screen telling participants to prepare for the recognition memory test. Participants were presented with 10 words sequentially on screen (one from each sentence) and asked to indicate whether or not they remembered hearing it: the letter "M" for "yes" and the letter "Z" for "no". This was reversed for left handed participants (N=1) such that the dominant hand always indicated a positive response. Target words remained on screen for five seconds, or until the participant responded. Participants were instructed to respond as quickly and accurately as possible. Order of target words in the recognition memory test was identical to the order of sentence presentation in to keep the time delay between presentation and their respective targets as consistent as possible.

Participants were permitted to take short breaks after each recognition memory test, before continuing with the next block of sentences. A single practice block was given prior to the commencement of the experiment to ensure participants were comfortable with the method. The experimenter was present for this time to answer questions.

2.2. Results

2.2.1. *Acoustic analyses of stimuli*

In order to validate the location of prosodic focus on the target words, for experimental and control sentences, target words were segmented and annotated based on inspection of the waveform and spectrogram in PRAAT [11]. For each target, duration, pitch (F0 mean, min, and max), and intensity (mean) values were extracted. Table 1 shows mean values for each of these measures as a function of condition.

Table 1. *Mean acoustic values for target words in each focus condition.*

Condition	NF	PR	ST	PS
Duration	305.971	433.85	297.121	408.743
f0_min	157.336	168.707	193.971	183.429
f0_mean	211.264	233.014	236.257	249.507
f0_max	326.979	420.386	355.807	434.229
int_mean	51.736	55.343	54.836	56.886

2.2.2. *Recognition accuracy scores*

Recognition accuracy scores were aggregated within condition, to yield a proportion of correct responses for each participant. Figure 1 displays the mean response accuracy as a function of condition. We used a within-subjects ANOVA with focus conditions as the independent variable for analysis.

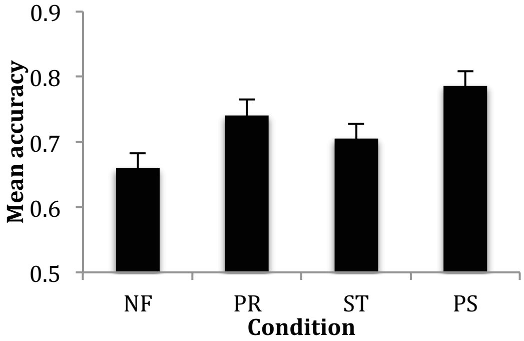


Figure 1. *Mean accuracy as a function of focus condition. Error bars represent standard error of the mean.*

The overall model was significant, F(3,31) = 16.69, p<.001. Words with focus (any type) were recognised more accurately than words without focus, F(1,31) = 31.53, p<.001. Words with prosodic focus were more likely to be recognised than words with no focus, F(1,31) = 16.39, p<.001. Words with both prosodic and syntactic focus were recognised more accurately than words in the syntactic focus alone condition, F(1,30) = 18.21, p<.001.

2.2.3. *Reaction time results*

Reaction times were aggregated to create a mean reaction time per condition dependent variable. We used a within subjects ANOVA with custom contrasts with condition as an independent variable. Figure 2 displays mean reaction time as a function of condition. The overall model was significant, F(3,31) = 11.36, p<.001. Words with focus (any type) were recognised significantly faster than words without focus, F(1,31) = 12.12, p=.001. Words with prosodic focus were recognised significantly faster than words in the no focus condition, F(1,31) = 5.12, p=.03. Words with some kind of syntactic focus were recognised significantly faster than words with prosodic focus alone, F(1,31) = 6.09, p=.03. Words with both prosodic and syntactic focus were recognised significantly faster than words with only syntactic focus, F(1,31) = 14.40, p=.001.

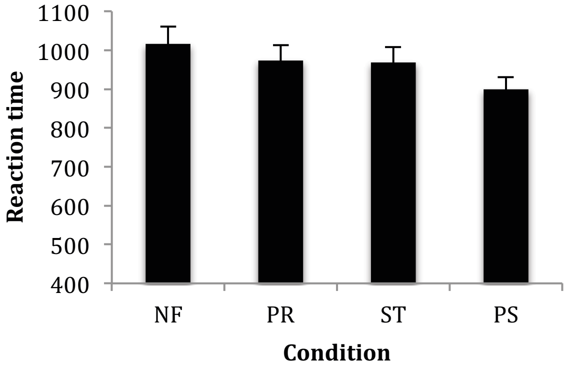


Figure 2. *Mean reaction times as a function of focus condition. Error bars represent standard error of the mean.*

2.2.4. *Comparing experimental and control sentences*

We further compared reaction time for words from experimental sentences to the control sentences. First we compared words that appeared as early targets (syntactic focus, prosodic + syntactic) with the early targets in the control sentences. Using a within subjects ANOVA with follow up Bonferroni adjusted pairwise comparisons, we showed the overall model was significant, F(2,31) = 18.19, p<.001, but crucially, both the experimental conditions were responded to significantly faster than the control targets, (M difference = 111.70, p=.002 for ST, and M difference = 178.99, p<.001 for PS). We then used a similar within subjects ANOVA to compare words that appeared as late targets (no focus, prosodic focus) with early control targets. The overall ANOVA was non significant, F(2,31) = 2.32, p=.11, however crucially, prosodically focused words were responded to significantly faster than late control targets (M difference = 52.41, p=.05).

2.3. Discussion

Our results showed that words with focus in English (whether conveyed by syntax or prosody) are more likely to be remembered, and are recognised faster than words without focus. Within the focus types, it seemed that words marked with some kind of prosodic focus received more of a facilitatory effect than words with syntactic focus. We also found an additive effect, such that words with both prosodic and syntactic focus were even more likely to be recognised, than words with either prosodic or syntactic focus alone. The reaction time results tell a somewhat different story, in that words with syntactic focus were recognised significantly faster than words with prosodic focus, but as with the recognition accuracy scores, there was an additive effect, such that words with both syntactic and prosodic focus were reacted to faster than words with a single focus type.

By comparing reaction times for target words from the control sentences with target words from the experimental sentences, we were able to determine that our findings were not simply an effect of regency; rather, word processing was facilitated by our manipulations of focus.

Our findings support previous work showing an additive effect for focus types. We also supported the hypothesis of a central search mechanism that seeks important information within an utterance to facilitate understanding [10].

3. Experiment 2

In Experiment 2, we investigated the underlying mechanisms for these processing advantages. We replicated Experiment 1, but changed the target words presented to participants, such that some were identical targets, and others were related either semantically, or phonologically.

3.1. Method

3.1.1. *Participants*

32 native speakers of Australian English were recruited for participation (M_{age} = 21.19, SD_{age} = 3.75).

3.1.2. *Stimuli and procedure*

The same sentences were used as Experiment 1. For 46 experimental sentences the same identical targets were used as above. For 47 sentences a phonologically related target word was selected (e.g. phonologically related target *shepherd* in place of the original *sheriff*). For the remaining 47 sentences a semantically related target was selected using the [12] database (e.g. semantically related target *tiger* in place of the original *lion*). The procedure was identical to Experiment 1.

3.2. Results

3.2.1. *Recognition accuracy scores*

As above, the recognition accuracy scores were aggregated to create a proportion correct dependent variable. We used a within subjects ANOVA with focus condition and target type as the independent variables.

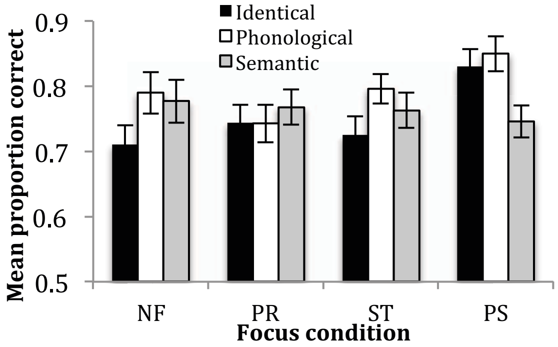


Figure 3. *The interaction between focus condition and target type. Error bars represent standard error of the mean.*

The main effect of focus condition was significant, F(3,31) = 4.51, p=.005, however the main effect of target type was not significant, F(2,31) = .99, p=.38. There was however, an interaction between focus condition and target type, F(6,31) = 3.56, p=.002 (displayed in Figure 3). The pattern of results for the identical targets mirrored Experiment 1. For semantic targets, response accuracy was consistent, independent of the focus condition, and for phonological targets, they were recognised more accurately, when in the prosodic + syntactic condition versus the other focus conditions.

3.2.2. Reaction time results

We used a within subjects ANOVA with focus condition and target types as the independent variables. The overall effect of focus condition was significant, F(3,31) = .014, as was the main effect of target type, F(2,31) = .022. There was no significant interaction between these two variables, F(6,31) = 1.70, p=.124. The effects of focus condition replicated those described above, therefore here we only explain the effect of target type. Overall, participants responded fastest to identical targets, followed by phonological targets, followed by semantic targets. Figure 4 shows the different target types as a function of focus condition.

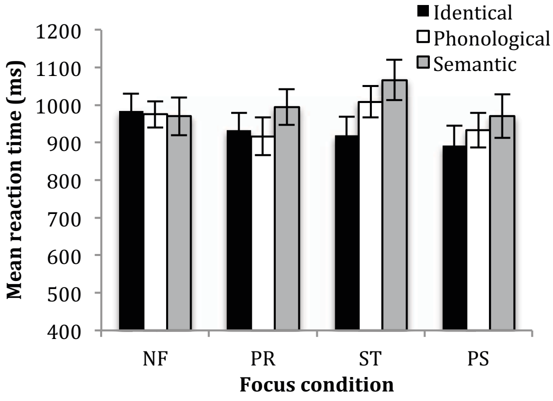


Figure 4. *Effect of focus condition and target type on reaction time. Error bars represent standard error of the mean.*

3.3. Discussion

The findings of Experiment 2, firstly replicate the effects of focus condition shown in Experiment 1. As for target type, when looking at the recognition memory results, there was no main effect of target type, but there was an interaction between target type and focus condition. For reaction time there was a main effect of target type, but no interaction. The findings together suggest that overall, identical targets are more likely to be recognised, and recognised faster, in a pattern consistent with Experiment 1. Phonological information was activated and responded to more accurately and faster in some focus conditions, but participants' ability to accurately respond to semantically related words did not seem to be related to which focus condition it was in. This suggests activation of a general semantic representation for the entire utterance, but that words with focus are attended to in greater detail via encoding of the specific word identity and phonological information.

4. General discussion

Taken together, these two experiments show that both focus types confer processing advantages in sentence processing in English and that this processing benefit is a specific memory trace of the target word itself, and its phonological information, rather than a general semantic representation.

This is the first study to our knowledge comparing the effect of syntactic and prosodic focus on processing in a single study in English. Consistent with [6], who showed an additive effect for semantic and prosodic focus processing, we also show an additive effect for syntactic and prosodic processing.

Our previous work in Korean also showed that while both focus types provided processing advantages, syntactic focus had more of an effect than prosodic focus on recognition memory [9]. Further cross-linguistic study is warranted.

5. Conclusions

These results contribute to a growing body of work assessing the role that focus plays in sentence processing. While both focus types do support efficient processing of words in English, it seems this effect is magnified when they work in concert. This lends weight to the hypothesis that it is a shared underlying mechanism that seeks new or important information within an utterance and directs attention to it.

6. Acknowledgements

This research was conducted with support from the ARC Centre of Excellence for the Dynamics of Language (project ID: CE140100041). The second author was also supported by a National Research Foundation of Korea Grant (2016S1A5B5A01025371).

7. References

[1] M. Krifka, "Basic notions of information structure," *Acta Linguistica Hungarica,* vol. 55, pp. 243-276, 2008.

[2] D. Bolinger, "Intonation across languages," *Universals of human language,* vol. 2, pp. 471-524, 1978.

[3] S. L. Birch, J. E. Albrecht, and J. L. Myers, "Syntactic focusing structures influence discourse processing," *Discourse Processes,* vol. 30, pp. 285-304, 2000.

[4] S. L. Birch and S. M. Garnsey, "The Effect of Focus on Memory for Words in Sentences," *Journal of Memory and Language,* vol. 34, pp. 232-267, 4// 1995.

[5] A. Cutler and D. J. Foss, "On the Role of Sentence Stress in Sentence Processing," *Language and Speech,* vol. 20, pp. 1-10, January 1, 1977 1977.

[6] E. Akker and A. Cutler, "Prosodic cues to semantic structure in native and nonnative listening," *Bilingualism: Language and Cognition,* vol. 6, pp. 81-96, 2003.

[7] A. Cutler, "Phoneme-monitoring reaction time as a function of preceding intonation contour," *Perception & Psychophysics,* vol. 20, pp. 55-60, 1976.

[8] Y.-C. Lee, B. Wang, S. Chen, M. Adda-Decker, A. Amelot, S. Nambu*, et al.*, "A crosslinguistic study of prosodic focus," *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing,* 2015.

[9] H. Kember, J. Choi, and A. Cutler, "Processing advantages for focused words in Korean," In J. Barnes, A. Brugos, S. Shattuck-Hufnagel, & N. Veilleux (Eds.), *Proceedings of Speech Prosody 2016* (pp. 702-705).

[10] A. Cutler and J. A. Fodor, "Semantic focus and sentence comprehension," *Cognition,* vol. 7, pp. 49-59, 1979.

[11] P. Boersma and D. Weenink. (2014, 2 January). *Praat: doing phonetics by computer (5.3.62 ed.).*

[12] D. L. Nelson, McEvoy, C. L., & Schreiber, T. A., "The University of South Florida word association, rhyme, and word fragment norms," http://www.usf.edu/FreeAssociation/, 1998.

Pitch accent type affects lexical activation in German: Evidence from eye tracking

Katharina Zahner, Muna Schönhuber, Janet Grijzenhout, Bettina Braun

Department of Linguistics, University of Konstanz

{katharina.zahner|muna.schoenhuber|janet.grijzenhout|bettina.braun}@uni-konstanz.de

Abstract

This visual world eye tracking study tests how f0 affects stress perception in online speech comprehension. The screen showed segmentally overlapping cohort pairs with different stress patterns (WSW/SWW) together with two distractors. In experimental trials, auditory stimuli referred to the WSW cohort member, which was presented with a medial-peak (L+H* L-%) or an early-peak pitch accent (H+L* L-%). Prior to segmental disambiguation, participants fixated the SWW stress competitor more when the WSW target was presented with an early-peak accent. Hence, the peak position affects lexical activation, such that pitch peaks preceding stressed syllables in WSW words temporarily activate SWW words.

Index Terms: eye tracking, stress, pitch accent type, German

1. Introduction

In languages with free stress (e.g., English, German or Dutch, cf. [1]), lexical stress distinguishes between lexical candidates (e.g., English *permit* vs. *permit*; underlining indicates lexical stress). Even when there is no exact stress-minimal pair, lexical stress reduces the number of lexical candidates [2, 3] and strongly guides lexical activation [4-7]. Cross-modal priming experiments with word fragments (cf. [5-7]) indicated that listeners are faster in recognizing words when the prime matched the stress pattern of the target (e.g., "octo" for "octopus/October", see [5]) than when it mismatched. Visual-world eye tracking studies further showed that suprasegmental information immediately modulates word recognition (cf. [4]): Dutch listeners use duration and intensity to resolve lexical competition prior to segmental disambiguation [4].

Lexical stress is an abstract notion at the word-level, which is acoustically cued by duration, intensity, and spectral energy. In German, stressed syllables are longer [8] and louder [8, 9] than unstressed ones, produced with increased vocal effort [10] and more peripheral vowel quality [11]. When accented, stressed syllables are additionally associated with a pitch movement which varies in regard to its position to the stressed syllable [12]. Many early studies have considered pitch a cue to lexical stress, with stressed syllables showing a higher f0 than unstressed ones (cf. [13, 14]). Yet, pitch movements are induced by phrase-level intonation, hence operating on a different linguistic level. Here, we investigate whether different pitch accent types (part of phrase-level phonology) affect the processing of metrical stress (word-level phonology) by studying lexical activation in German.

Theoretically speaking, it may seem odd to expect that intonation could have any effect on lexical access in an intonation language (in which pitch is not lexically contrastive); yet, previous research demonstrated that lexical processing is indeed affected by intonation. For instance, an unfamiliar intonation contour leads to slower lexical access in Dutch [15], listeners rely on the pitch contour to decide between lexical candidates in Italian, English and French [16-18] and German listeners are slower in processing isolated SW words whose f0-contours are taken from WS words [19].

More importantly, recent studies on German showed that phrase-level intonation, i.e. pitch accent type, influences infants' and adults' perception of metrical stress. (Note that in spoken communication, the choice of accent type is influenced by pragmatic considerations: While early-peak accents signal information that is accessible or predictable for listeners [20, 21], medial-peak accents are associated with information that is newly introduced into the discourse [21].) Specifically, [22] showed that German nine-month-olds rely on the position of the pitch peak as a cue to stress, such that only high-pitched stressed syllables are used for the extraction of trochees from fluent speech. High pitch even outweighs other stress cues [23]: Infants treat high-pitched unstressed syllables erroneously as stressed, taking them as the strong (stressed) element when segmenting speech. Furthermore, in an offline-paradigm, German adults have been demonstrated to make more errors in identifying the stressed syllable when the word is produced with a pitch accent that renders the stressed syllable low-pitched [24]. However, these studies do not answer the question whether and how pitch information influences German adults in online word recognition.

Taken together, we predict that phrase-level intonation modulates lexical access in online speech comprehension. More specifically, high-pitched but unstressed syllables in WSW words are expected to be perceived as stressed, leading to the temporary activation of cohort competitors with an "opposite" stress pattern, i.e. a SWW word-prosodic structure.

2. Experiment

We investigate how pitch accent type affects lexical activation in German adults by using eye tracking (visual world) with four printed words on screen [25], a paradigm that is sensitive to phonetic/phonological differences. We particularly focus on two kinds of rising-falling accents that differ in the alignment of the tonal movement with the stressed syllable: In early-peak accents, the pitch peak precedes the stressed syllable (H+L*), while in medial-peak accents (L+H*) the stressed syllable and the pitch peak coincide. Similar to [4], the screen showed two written representations of trisyllabic cohort competitors that differ in the position of stress (e.g., WSW "Libelle" 'dragonfly' and SWW "Libero" 'sweeper'). We tested whether high-pitched unstressed syllables in WSW words, i.e., syllables on which a pitch peak is realized, but which do not carry lexical stress (as in early-peak accents, H+L*), prompt adults to perceive them as stressed, thus activating the cohort member with initial stress (SWW).

2.1. Methods

2.1.1. *Participants*

Forty-eight German native speakers (38 female, ∅ = 23.9 years, SD = 3.1 years) with normal or corrected-to-normal vision and unimpaired hearing participated for a small fee.

2.1.2. *Materials*

Sixty-four trisyllabic cohort pairs that differed in the position of stress (WSW vs. SWW words) were selected. The cohort pairs were segmentally identical up to at least the onset consonant of the second syllable, e.g., “Libelle” [li'bɛlə] vs. “Libero” ['libəʁo]. Cohort members were matched for lexical frequency and number of characters across groups. For each cohort pair, we selected two semantically and phonologically unrelated distractors with comparable number of characters and frequency to be presented on screen. Distractors were SWW-, WSW- or WWS words (one third in each pattern). For 32 of the 64 cohort pairs, the auditory target was one of the cohort members (in 16 experimental trials the WSW word, in 16 distractor trials the SWW word), in the other 32 trials the auditory target referred to one of the unrelated words (filler trials). For instance, in experimental trials, the WSW cohort member “Libelle” was presented as auditory target and the SWW cohort member “Libero” was the stress competitor.

The auditory targets were embedded in a semantically non-constraining carrier sentence (e.g., “Bitte klicke Libelle an”, ‘Please click on dragonfly’). A female native speaker of Standard German recorded the sentences in a sound-attenuated cabin in the PhonLab at the University of Konstanz (44.1kHz, 16Bit). The auditory targets for experimental (WSW as auditory target) and distractor trials (SWW as auditory target) were produced in two intonation conditions: with a medial-peak (L+H*) and an early-peak accent (H+L*, see Figure 1). The recordings in the early- and medial-peak condition were matched along a number of acoustic variables; see Table 1 for measurements in experimental trials. Similarly, half of the auditory targets for the fillers were recorded with a medial-peak, half with an early-peak accent. To reduce distal prosodic context effects [26], stimuli were spliced into a carrier sentence “Bitte klicke”. Four different carriers (same across conditions) were used for targets starting with different vowels ([a] vs. [e]), the consonant [m] or any other consonant to avoid co-articulation. Thus, the carrier was identical for each cohort pair in both intonation conditions. The cross-spliced stimuli were rated to sound natural and splicing was not noticeable. Words in the carrier were not accented (see Figure 1) to avoid metrical expectations based on the preceding f0-contour [26].

2.1.3. *Procedure*

Participants were tested individually in the PhonLab at the University of Konstanz. They were seated approximately 70cm in front of a LCD screen (37.5cm x 30cm). The Desktop Mount was used with a head support. The dominant eye was calibrated (pupil and corneal reflection) in an automatic procedure, using the SR Eyelink 1000 Plus tracking system at a sampling rate of 250Hz.

The experiment consisted of 64 trials, 16 experimental trials (WSW cohort member as auditory target), 16 distractor trials (SWW cohort member as auditory target), and all 32 filler trials (unrelated distractor as auditory target). In experimental and distractor trials, intonation condition (medial- vs. early-peak) was distributed in a Latin-Square Design, i.e., each subject heard both intonation conditions, but each item in only one of the conditions. Eight experimental lists were created, pseudo-randomizing the order of trials such that each half contained the same number of cohort and distractor trials, with the constraint of the experimental item (WSW) being at most the third item of the same condition in a row. Each list started with five filler and two distractor trials to familiarize participants with the task and voice. Participants were assigned randomly to one of the experimental lists (six participants per list, eight items in each condition).

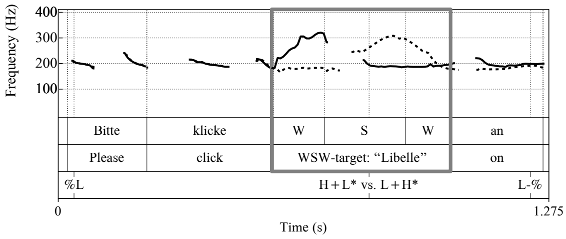


Figure 1: *F0-contour of two experimental trials (medial-peak (dashed); early-peak condition (solid)).*

Table 1. *Acoustic realization means (and standard deviations) of WSW targets in two intonation conditions (experimental trials).*

Acoustic variable	Medial-Peak Condition	Early-Peak Condition
F0-excursion of accentual movement in st	Rise: 8.36 (0.60)	Fall: 8.43 (0.67)
Slope of accentual movement in Hz/sec	51.6 (11.7)	54.8 (9.7)
Duration of first syllable in ms	143 (34)	146 (34)
Duration of second syllable in ms	214 (48)	226 (48)
Duration of third syllable in ms	164 (34)	157 (35)
H1*-A3* ratio ([10]) in middle of first vowel in dB	27.0 (10.8)	23.2 (9.7)
H1*-A3* ratio in middle of second vowel in dB	30.8 (7.2)	31.8 (6.3)
H1*-A3* ratio in middle of third vowel in dB	27.7 (5.5)	28.7 (4.5)

Every trial started with a fixation cross that was centered on screen and displayed until participants clicked on it. Upon clicking, the four words appeared on screen (Times New Roman Font, size 20). They were presented in the outer third of the four quadrants of the screen (to avoid peripheral looking) and framed by a rectangular box. The position of the cohort members and distractors were randomized on screen, such that in all lists the auditory target occurred equally frequent in all four possible positions. The carrier phrase started 2000ms after the words occurred on screen, leaving a preview of 1426ms on average. Participants were instructed to click on the word named in the auditory stimulus as fast as possible. Auditory stimuli were presented via headphones (Beyerdynamic DT-990 Pro, 250 OHM) at comfortable loudness. Every fifth trial, a drift correction was initiated. After half of the trials, there was an optional pause. The total duration of the experiment was approximately 15 minutes.

Table of Contents for this manuscript

2.2. Results

For reasons of space, we only report on experimental trials here. In these trials, participants clicked on the auditory target (WSW) in 98.3% of all trials; there was no effect of intonation condition on error rates or reaction times (both p > 0.3).

The eye tracking data were extracted in 4ms steps and coded into saccades, fixations, and blinks (default settings for normal saccade sensitivity in the EyeLink 1000 software); only fixations were further processed. For experimental trials, fixations were automatically labeled as being directed to the target (e.g., "Libelle"), the stress competitor (e.g., "Libero"), or to the distractors if they fell within a square of 200x200 pixels around the respective word. Figure 2 shows the evolution of fixations in experimental trials to the words on screen in the two intonation conditions. The dashed vertical lines indicate the acoustic landmarks of the auditory stimuli.

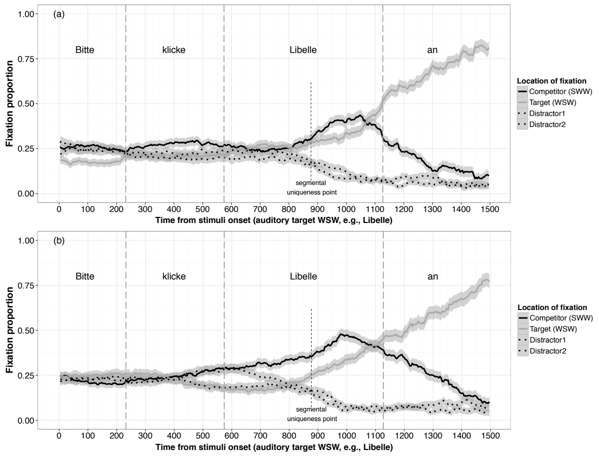


Figure 2: *Evolution of fixations to competitor, target and the two distractors in the medial-peak condition (a) and the early-peak condition (b).*

For statistical analysis of the SWW competitor activation in experimental trials, empirical logits of fixations to the competitor were calculated (ratio of the fixations to the competitor divided by the fixations directed to the three other objects or somewhere else [27]). We fitted a linear mixed effects regression model with *intonation condition* (medial-peak vs. early-peak) as fixed factor and *participants* and *items* as crossed random factors [28], allowing for adjustments of intercepts and slopes [29]. P-values were obtained using the Satterthwaite's approximation (R package lmerTest [30]).

Results showed an effect of *intonation condition* on the fixations to the stress competitor (SWW) in the window in which suprasegmental information differed across conditions but segmental information did not distinguish the lexical candidates. During segmental ambiguity, participants fixated the SWW stress competitor more often when the WSW target was realized with an early-peak (average elogs -1.14), compared to a medial-peak accent (average elogs -1.67, ß = 0.5 [0.02;1.04], SE = 0.26, t = 2.05, p = 0.04). This difference in fixations to the SWW competitor is illustrated more directly in Figure 3. Figures 2 and 3 further suggest differences in fixations to the stress competitor already before information on the auditory target became available. This difference is not significant, however (p = 0.24). Fixations to the target also differed as a function of intonation condition: There were significantly more fixations to the WSW target in the medial-peak, compared to the early-peak condition, but only during the processing of the early part of the first syllable (720ms-800ms; ß = 0.5 [0.05; 0.93], SE = 0.22, t = 2.22, p = 0.03).

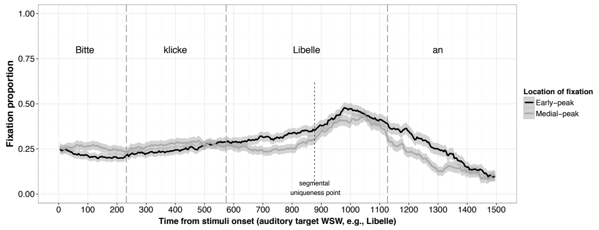


Figure 3: *Fixations to SWW stress competitor in experimental trials (WSW auditory target) in the two intonation conditions.*

2.3. Discussion

Our results reveal intonational interference during lexical activation: There were more fixations to the stress competitor (SWW) – and consequently fewer fixations to the target (WSW) – when the WSW target was presented with an early-peak accent (H+L*) than when presented with a medial-peak accent (L+H*). This shows that high-pitched unstressed syllables temporarily activate competitors with initial stress.

Note that both medial-peak (L+H*) and early-peak accent (H+L*) naturally occur in German. The carrier sentence used in the experiment might favor a medial-peak accent, since the objects to be clicked on were new referents [21]. Yet, the repeated mention of the same carrier evokes a notion of accessibility of the objects as a whole, which makes an early-peak realization equally pragmatically appropriate [20].

3. General discussion and conclusion

Overall, we find that pitch accent type influences lexical activation: German adults inadvertently activate words with the wrong stress pattern (SWW) when presented with WSW words in which the pitch peak precedes the stressed syllable. During segmental ambiguity, high-pitched unstressed syllables are more often interpreted as stressed than low-pitched unstressed syllables. Note that in both intonation conditions, the prosodic stress cues suggest that the second syllable is stressed. The cue that varies across intonation conditions is the position of the pitch peak. High pitch hence seems to be a relevant cue to metrical stress for German adults. Even though theoretically different from word-level stress, pitch accent type affects the processing of metrical stress. Here, we demonstrate that pitch information influences German listeners in online word recognition, not only in offline metalinguistic stress judgments [24]. We further show that the association between high pitch and metrical stress observed in German infants [22, 23] is also found in German adults.

How can we explain that pitch accent type influences lexical processing in that way? It is striking that adults interpret high-pitched unstressed syllables erroneously as stressed, since due to phrase-level intonation this strategy is not profitable in all cases. In spoken communication, using f0 as a cue to stress might even lead to higher processing costs, as there is more lexical competition. We see three explanations why high-pitched syllables play a role in the online processing of metrical stress: First, high-pitched syllables are perceptually salient [31]. Listeners might thus equate perceived acoustical prominence with metrical prominence. Second, medial-peak

Table of Contents for this manuscript

are more frequent in German spontaneous speech than early-peak accents [32]. Conceivably, the frequent encounter of high-pitched stressed syllables might make the pitch peak a cue to stress. Finally, the findings in the framework of the iambic-trochaic law might shed light on the observed pattern of results: [33], for instance, showed that adults tend to group pairs of syllables alternating in pitch with initial prominence (SW). Future research with other pitch accent types and/or other languages is needed to better understand the mechanism.

Here, we show that high pitch is a cue to stress in German adults' online processing, despite the fact that pitch accent type is part of phrase- and not word-level phonology. Hence, in speech processing f0 is more intertwined with other stress cues than in phonological theory. Consequently, this finding poses questions for spoken word recognition models (e.g., Shortlist [34], Shortlist B [35], FUL [36]) that currently do not account for utterance-level intonation.

4. References

[1] U. Domahs, I. Plag, and R. Carroll, "Word stress assignment in German, English and Dutch: Quantity-sensitivity and extrametricality revisited," *The Journal of Comparative Germanic Linguistics,* vol. 17, pp. 59-96, 2014.

[2] A. Cutler, *Native listening: Language experience and the recognition of spoken words*. Cambridge, Mass. [u.a.]: MIT Press, 2012.

[3] A. Cutler and D. Pasveer, "Explaining cross-linguistic differences in effects of lexical stress on spoken-word recognition," *Proceedings of 3rd Internatinal Conference on Speech Prosody*, Dresden, 2006.

[4] E. Reinisch, A. Jesse, and J. M. McQueen, "Early use of phonetic information in spoken word recognition: Lexical stress drives eye movements immediately," *The Quarterly Journal of Experimental Psychology,* vol. 63, pp. 772-783, 2010.

[5] M. Koster, W. van Donselaar, and A. Cutler, "Exploring the role of lexical stress in lexical recognition," *The Quarterly Journal of Experimental Psychology,* vol. 58, pp. 251-274, 2005.

[6] C. K. Friedrich, S. A. Kotz, A. D. Friederici, and T. C. Gunter, "ERPs reflect lexical identification in word fragment priming," *Journal of Cognitive Neuroscience,* vol. 16, pp. 541-552, 2004.

[7] N. Cooper, R. Wales, and A. Cutler, "Constraints of lexical stress on lexical access in English: Evidence from native and non-native listeners," *Language and Speech,* vol. 45, pp. 207-228, 2002.

[8] M. Jessen, K. Marasek, and K. Claßen, "Acoustic correlates of word stress and the tense/lax opposition in the vowel system of German.," *Proceedings of the 13th International Congress of the Phonetic Sciences,* Stockholm, 1995.

[9] G. Dogil, "Phonetic correlates of word stress," *Arbeitspapiere des Instituts für Maschinelle Sprachverarbeitung (Univ. Stuttgart),* vol. 2, pp. 1-60, 1995.

[10] C. Mooshammer, "Acoustic and laryngographic measures of the laryngeal reflexes of linguistic prominence and vocal effort in German," *Journal of Acoustical Society of America,* vol. 127, pp. 1047-1058, 2010.

[11] P. Delattre, "An acoustic and articulatory study of vowel reduction in four languages," *International Review of Applied Linguistics and Language Teaching (IRAL),* vol. 7, pp. 294-325, 1969.

[12] J. B. Pierrehumbert, "The Phonology and Phonetics of English intonation," Massachusetts Institute of Technology, Dept. of Linguistics and Philosophy, Bloomington, 1980.

[13] D. B. Fry, "Experiments in the perception of stress," *Language and Speech,* vol. 1, p. 126, 1958.

[14] I. Lehiste, *Suprasegmentals*. Cambridge, Mass. [u.a.]: MIT Press, 1970.

[15] B. Braun, A. Dainora, and M. Ernestus, "An unfamiliar intonation contour slows down online speech comprehension," *Language and Cognitive Processes,* vol. 26, pp. 350-375, 2011.

[16] M. D'Imperio, C. Petrone, and N. Nguyen, "Effects of tonal alignment on lexical identification in Italian," in *Tones and Tunes*, C. Gussenhoven and T. Riad, Eds., Berlin: Mouton de Gruyter, 2007, pp. 79-106.

[17] D. R. Ladd and A. Schepman, "“Sagging transitions” between high pitch accents in English: Experimental evidence," *Journal of Phonetics,* vol. 31, pp. 81-112, 2003.

[18] P. Welby, "The role of early fundamental frequency rises and elbows in French word segmentation," *Speech Communication,* vol. 49, pp. 28-48, 2007.

[19] C. K. Friedrich, K. Alter, and S. A. Kotz, "An electrophysiological response to different pitch contours in words," *Neuroreport,* vol. 12, pp. 3189-3191, 2001.

[20] S. Baumann and M. Grice, "The intonation of accessibility," *Journal of Pragmatics,* vol. 38, pp. 1636-1657, 2006.

[21] K. Kohler, "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.

[22] K. Zahner, M. Schönhuber, and B. Braun, "The limits of metrical segmentation: Intonation modulates infants' extraction of embedded trochees," *Journal of Child Language,* pp. 1-27, 2015.

[23] K. Zahner, M. Schönhuber, J. Grijzenhout, and B. Braun, "High pitch signals word onsets for German 9-month-olds: Evidence for a pitch-based segmentation strategy," Presented at *Tone and Intonation in Europe (TIE)*, Canterbury, UK, 2016.

[24] S. Egger, "The impact of pitch accents on the identification of word stress in German," MA, Department of Linguistics, University of Konstanz, Konstanz, 2015.

[25] J. M. McQueen and M. Viebahn, "Tracking recognition of spoken words by tracking looks to printed words," *The Quarterly Journal of Experimental Psychology,* vol. 60, pp. 661-671, 2007.

[26] M. Brown, A. P. Salverda, L. C. Dilley, and M. K. Tanenhaus, "Metrical expectations from preceding prosody influence perception of lexical stress," *Journal of Experimental Psychology: Human Perception and Performance,* vol. 41, pp. 306-323, 2015.

[27] D. J. Barr, T. M. Gann, and R. S. Pierce, "Anticipatory baseline effects and information integration in visual world studies," *Acta Psychologica,* vol. 137, pp. 201-207, 2011.

[28] R. H. Baayen, *Analyzing linguistic data: A practical introduction to statistics using R*. Cambridge [u.a.]: Cambridge Univ. Press, 2008.

[29] D. J. Barr, R. Levy, C. Scheepers, and H. Tily, "Random-effects structure for confirmatory hypothesis testing: Keep it maximal," *Journal of Memory and Language,* vol. 36, pp. 255-278, 2013.

[30] A. Kuznetsova, P. B. Brockhoff, and R. H. B. Christensen. (2013). *lmerTest: Tests for random and fixed effects for linear mixed effect models*. Available: http://cran.r-project.org/web/packages/lmerTest/index.html.

[31] S. Baumann and C. Röhr, "The perceptual prominence of pitch accent types in German," *Proceedings of the 18th International Congress of the Phonetic Sciences*, Glasgow, UK, 2015.

[32] B. Peters, K. Kohler, and T. Wesener, "Melodische Satzakzentmuster in prosodischen Phrasen deutscher Spontansprache - Statistische Verteilung und sprachliche Funktion," in *Prosodic Structures in German Spontaneous Speech (AIPUK 35a)*, K. Kohler, F. Kleber, and B. Peters, Eds., Kiel: IPDS, 2005, pp. 185-201.

[33] R. A. H. Bion, S. Benavides-Varela, and M. Nespor, "Acoustic markers of prominence influence infants' and adults' segmentation of speech sequences," *Language and Speech,* vol. 54, pp. 123-140, 2011.

[34] D. Norris, "Shortlist: A connectionist model of continuous speech recognition," *Cognition,* vol. 52, pp. 189-234, 1994.

[35] D. Norris and J. M. McQueen, "Shortlist B: A Bayesian model of continuous speech recognition," *Psychological Review,* vol. 115, pp. 357-395, 2008.

[36] A. Lahiri and H. Reetz, "Underspecified recognition," in *Labphon*. vol. 7, C. Gussenhoven, N. Werner, and T. Rietveld, Eds., Berlin: Mouton, 2002, pp. 637-676.

Evidence and Intonational Contours: An Experimental Approach to Meaning in Intonation

Byron Ahn[1], *Stefanie Shattuck-Hufnagel*[2], *Nanette Veilleux*[3]

[1]Princeton University, USA
[2]Massachusetts Institute of Technology, USA
[3]Simmons College, USA

bta@princeton.edu, sshuf@mit.edu, veilleux@simmons.edu

Abstract

One proposed function of prosody is conveying speakers' stances on the evidence for their statements and assessments of listeners' beliefs. Testing this is challenging – specifying evidential status is difficult, and speakers vary intonationally for a given context. A novel task of reading lines from comic strips elicits relatively consistent intonation, suggesting both the method's usefulness, and the efficacy of speaker beliefs in governing prosodic contours. Preliminary results suggest H* accents are used when speakers believe they have evidence the hearer lacks, and L* accents for the flipped situation.

Index Terms: speech prosody, intonational meaning, pragmatics, semantics, experimental method

1. Introduction and Background

It is well-established that in languages like English, the meaning of a spoken utterance is signalled by the morpho-syntax (i.e., the segmentally-specified words/morphemes and syntactic structure of the sentence), and by the prosody (i.e., the grouping and prominence patterns conveyed by, e.g., the suprasegmental intonational tones/contours). The difficulty of designing rigorous experiments and defining appropriate units has hindered our understanding of the contribution of intonation to meaning. This paper presents an initial attempt at investigating the connection between formal models of meaning (semantics and pragmatics) and formal models of prosody, with the ultimate goal of creating a model that correctly predicts the appropriate prosodic contour(s) for any appropriately defined linguistic context. In particular, we test the hypothesis that the speaker's beliefs about the status of the evidence about a declaration influences the choice of intonational contour for the statement.

While past research by prosodic theorists have attempted to link elements of the prosodic inventory with meaning, most have found that their claims are limited and have been difficult to generalize. Intonationists have hypothesized mapping accent type to meanings on the basis of mutual beliefs (e.g., [1], [2]), but subsequent experimental testing has shown that these initial hypotheses require further elaboration. For example, the Low-High accent L+H* has been claimed to occur on words which explicitly contrast to an alternative. However, the distribution of L+H* accents does not fully support this prediction (e.g., [3]). We hypothesize this problem arises in part from the specificity and precision of the hypotheses. That is, accents have meanings, but their meanings have been described narrowly and incompletely. Further, we propose that advances in pragmatic and semantic theory, describing the aspects of meaning that contribute to discourse structure and how those dynamics influence the particular forms of sentences, can illuminate these issues. For example, some languages use "discourse particles" or "evidential markers" —segmental morphemes or words that indicate particular discourse structures—raising the question of how other languages (e.g. English) represent this conversational information. Recent research suggests that some of these discourse dynamics (i.e., who has what information) may be expressed intonationally in English (e.g., [4], [5]).

While the semantics/pragmatics literature suggests that intonation can signal important information ([6]), these works also tend to focus on a narrow set of prosodic features (e.g., edge tones). This approach has uncovered important findings, but has not explored potentially important generalizations about what types of meaning are carried by which intonational elements (pitch accent types and combinations of pitch accents with edge tones, i.e., 'tunes'). While some broader work on intonational meaning has been carried out ([1], [7], overview in [6]), many claims are yet to be rigorously tested experimentally, and few investigations have been framed in terms of contemporary theories of evidentiality or discourse pragmatics.

This project develops a new experimental paradigm, in which variables identified by recent advances in pragmatic theories are manipulated experimentally, building on earlier work to address the question of how meaning relates to intonation. Specifically, we bridge the conversational space between speakers and their interlocutors with an elicitation experiment with conditions defined in terms of sourcehood and evidentiality (cf. [4], [5]). Using this paradigm, we have found promising systematic results in prosodic behavior that suggest the semantic/pragmatic factors we manipulate play a role in determining the speaker's choice of intonational contour.

This approach will be useful for future work investigating more abstract questions about precisely how intonation carries particular meanings (e.g., tones vs. tunes; cf. [7], [1]). In addition, our results raise issues about the representational nature of discourse information in languages (like English) where such information is marked via intonation. As noted above, in many other languages, these pragmatic meanings are represented with segmental words/morphemes; is such information also syntactically represented in English-type languages? A standard view is that any pairing of phonetic/phonological information with semantic/pragmatic meaning is mediated by syntax ([8]:p.1), so that semantics/pragmatics cannot influence phonetics/phonology directly. If intonational meaning shares this characteristic, one would expect *syntactic* difference between sentences like "Are you tired↑ or sad↓?" (L* H- H* L-L%) and "Are you tired↑ or sad↑?" (L* H- L* H-H%), as in, e.g., [9], [10], [11]. In this way, our finding that the interpretive significance of intonation may track sourcehood/evidentiality is consistent with the view that (some) intonational tones/contours may be abstractly represented in the syntactic structure (also argued by, e.g., [12]). If this hypothesis is confirmed more broadly, it would mean that intonational work is not immune to a recurrent finding in linguistics: adequate work in any one subfield requires deep understanding of all the others.

2. Methodology

Production experiments are notoriously difficult to control. In the authors' experience (and in subject self-report) speakers often adopt unintended roles or simply begin to read prompts without any particular communicative intent. For this reason, laboratory speech prompts and perception studies, rather than production experiments, predominate in the literature, or physiological methods like eye-trackers are used to ascertain the perceptual impact. Clearly this poses difficulties for researchers intending to map prosodic production onto function [13]. In this work, we devise a novel production task using comic strips to more directly explore this mapping.

We aim to explore the role of evidence present to both the Speaker and Hearer in guiding the development of the conversational 'common ground', as reflected in the intonational choices of the Speaker. Our chosen framework predicts that the choice between different prosodic contours will be influenced by who the Speaker believes holds evidence for a proposition and in the strength of this evidence. In particular, we conclude that the edge tones will vary in direction (rising for questions, falling for confirmations) and pitch accents will vary in type (High, Low or bitonal combinations such as H*, L*, L+H* in MAE_ToBI notation; [14]) when subjects/speakers respond to conditions that vary the evidence that (the speaker believes) the hearer has.

In other words, this research seeks to experimentally determine whether there is a link between degrees of evidence and shared beliefs, on the one hand, and the implementation of the pitch accent and boundary tone, on the other. Manipulating the variables identified by Gunlogson and Northrup [4, 5], we have devised five basic semantic/pragmatic contexts for a spoken interaction between S (the speaker) and H (the hearer), where each condition controls what S and H know to be true and what S reasonably believes about what H believes to be true. Using controlled narrative conditions, we manipulate what evidence the speaker (S) believes the hearer (H) has about whether a proposition is true (e.g., H is a reliable source, H has direct, indirect, no or contradictory evidence; see Table 1). The five conditions create a partially ordered list, ranging from expected confirmation of shared beliefs (A) through declaring new information (D), to contradiction (E).

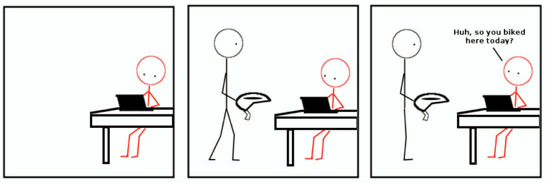
The novel methodology developed to address these issues uses dialogs to represent each of the five conditions (A-E in Table 1) for two different scenarios (raining and biking). The circumstances illustrated in 3-panel comic strips (See Figure 1) described the specific context. Ten subjects were recruited from a sample of convenience from the adult native English speaking population at the home institution of one of the authors. Subjects were instructed to read the comic strip silently, associating themselves with the character in the last panel (drawn in red) and then read that character's dialog aloud. Each of two scenarios, with prompts in all of the five contexts, were repeated a total of 3 times in non-consecutive but otherwise random order, for a total of 300 recorded responses. The analysis presented here covers Conditions A, B, and D (180 responses).

Our hypothesis is that the use of boundary (final) tones will be consistent with published accounts ([4]). As for pitch accents, literature does not (as far as we know) make any specific predictions about the relationship between pitch accents and evidentiality/sourcehood. We therefore cautiously hypothesize that the pitch accent of target words (e.g., "rain" or "bike" in the two scenarios) may depend on condition, as well.

Table 1: *Example of five evidentiality/sourcehood-based contexts and the relationship of each to the status of the Speaker's beliefs about the Hearer's belief concerning the proposition.*

Condition	Description of Evidence/Sourcehood
A	S asserts a proposition P, deferring to H as a reliable source of information, believing that H has direct evidence that P is true.
B	S asserts a proposition P, believing that H also has direct evidence that P is true.
C	S asserts a proposition P, believing that H has indirect evidence suggesting that P is true.
D	S asserts a proposition P, believing that H has no evidence about P's truth.
E	S asserts a proposition P, believing that H has a indirect evidence that contradicts P.

Figure 1: *An example of one prompt for Scenario A. The dialog bubble reads "Huh, so you biked here today?"*



3. Results

The productions of ten participants were labeled by three experienced annotators (the authors) using MAE_ToBI (e.g., [14]). Agreement between at least 2 of the 3 labels for a given utterance was taken as the correct label; where three different labels had been generated, consensus labels were determined through discussion. This resolved issues for all but one speaker, in whose productions the target word was deaccented at roughly twice the rate of other subjects. That speaker's responses were judged to be outside of the dialect described by MAE_ToBI, and their responses were removed from the analysis. Another 15 of the 162 responses where the target word was deaccented were also discarded (This was more frequent in the "bike" scenarios where the personal pronouns 'I' or 'she' competed with the verb for accentuation). Our Phase I analysis focused on contexts A, B and D for the remaining 9 subjects.

Notably, in the 147 recordings analyzed in this Phase I analysis, labellers agreed upon labels for pitch accents in 137 tokens (93%), and labels for boundary tones in 140 tokens (95%). When considering the entire contour, labellers agreed upon labels for both pitch accents and boundary tones in 131 tokens (89%). Moreover, results of this Phase I analysis reveal strong trends in production, indicating that the manipulated contextual variables (determined by discourse-structural notions of sourcehood/evidentiality) are closely related to the meaningful contributions of intonational tones/contours.

Examining pitch accents first, Tables 3 and 4 reveal strong correlations between discourse-context condition and pitch ac-

Table 2: *Example of five contexts and a specific scenario to elicit prosodic productions.*

Condition (from Table 1)	**Description of Scenario** (Biking version)
A: S asserts a proposition P deferring to H as a reliable source of information, believing that H has direct evidence that P is true.	H walks into the room, carrying a bike helmet. S says to H “Huh, so you biked here today?”
B: S asserts a proposition P, believing that H has direct evidence that P is true.	Persons X, S and H are co-workers. Person X rides by S and H on a bicycle. S says to H “Huh, so she biked here today?”
C: S asserts a proposition P, believing that H has indirect evidence suggesting that P is true.	S and H are co-workers. H sees S walking in with a bike helmet. Later, when talking about getting home, S says to H “I biked here today, you know”
D: S asserts a proposition P, believing that H has no evidence about P’s truth.	S and H are co-workers. H asks S how they got to work. S says “I biked here today.”
E: S asserts a proposition P, believing that H has a belief that contradicts P.	S and H are co-workers. S has a cast on her leg and S says to H “Guess what! I biked here today.”

cent choice. In both Conditions A and B, speakers used L-type (including L* and L*+H) accents most frequently (A: 37/46, 80%; B: 32/49, 65%), and in Condition D, speakers always used H-type (including H*, !H*, and all bitonal variants) accents.

Examining the contours in more detail, results that take boundary tone into account reveal that the preferred contour for A (30/46, 65%) uses an L* H-H% contour, the classic contour for a Yes-No question in American English; at the same time, the preferred contour for D (30/52, 58%) is H* L-L%, the classic contour for a neutral declarative statement. These results are given in Tables 5 and 6. (Note that in Table 6, the labels ‘H*’ and ‘L*’ do not include bitonal variants.)

Table 3: *Distribution of specific L and H-type accents across speakers for Conditions A, B, and D. Shading indicates the most common accent for the condition.*

Pitch Accent	**A**	**B**	**D**
H*	4	7	**35**
L+H*	5	10	5
!H*	0	0	1
H+!H*	0	0	11
L*	**32**	**29**	0
L*+H	5	3	0

Table 4: *Distribution of L-type and H-type accents across speakers for Conditions A, B, and D. Note the progression of a greater prevalence of H-type accents as S’s belief about H’s evidence for the proposition increases. Shading indicates the most common accent-type for the condition.*

Pitch Accent type	**A**	**B**	**D**
H-type	9	17	**52**
L-type	**37**	**32**	0

Table 5: *Distribution of contours with both L and H-type accents for four boundary tones (!H merged with H in all cases), across speakers for Conditions A, B, & D. Shading indicates the most common contour for the condition.*

Pitch Accent type	**Boundary Tone**	**A**	**B**	**D**
H-type	L-L%	9	15	**43**
	H-L%		1	6
	L-H%			3
	H-H%		1	
L-type	L-L%		1	
	H-L%	4	1	
	L-H%			
	H-H%	**33**	**30**	

Table 6: *Distribution of contours across speakers for Conditions A, B, & D, organized by boundary tone. ‘H*’ and ‘L*’ exclude bitonals. Shading indicates most common contour.*

Boundary Tone	**Pitch Accent**	**A**	**B**	**D**
L-L%	H*	4	7	**30**
	Other pitch accents	5	9	13
H-H%	L*	**30**	**28**	0
	Other pitch accents	3	3	0
Other boundary tones	Any pitch accent	4	2	9

Speakers also produced B with a preference for a L* H-H% contour (28/49, 57%) but this preference was less pronounced. That is, for the B condition, H-type accents appeared in about 35% of tokens (17/49), with 59% of those (10/17) appearing as L+H*. Although the data should be considered preliminary, this suggests an evolution with respect to the alignment along a scale of certainty about S’s beliefs about H’s evidence for the proposition at hand. Moreover, it is our impression that this shared ToBI label for conditions A and B hides a reliable difference in scaling; further analysis at the acoustic level will be required to test this hypothesis.

These results are remarkable in their high level of consistency for a given context (A, B or D) across speakers and (more notably) *within* speakers across repetitions (which were not elicited consecutively). Individual speakers produced identical pitch accents and boundary tones across the entire triad of repetitions (the 3 non-consecutive repetitions of the prompt, for an individual speaker) 19 out of 54 times (2 scenarios, 3 contexts, 9 speakers). Accents were identical in 24 triads – even when counting, e.g., !H* and H* as distinct; collapsing L-type and H-type accents raises this within-speaker consistency to 36 of 54 triads. Boundary tones were also produced with a high degree of consistency for entire triads; speakers produced identical boundary tones within a triad in 35 of 54 cases.

In sum, each context (defined by manipulating evidentiality-based variables) elicited consistent intonational contours from individuals. This intra-speaker consistency is valuable because a given speaker, with a consistent understanding of both the evidence and the speaker's/hearer's relationship with that evidence, produces a consistent intonational contour under these experimental conditions. (Across speakers, there may be different ways of understanding the pragmatic context, allowing for the observed limited variation in intonation.) These findings support earlier proposals from formal pragmatic research that evidence strength and reliability variables impact a speaker's intonational choices ([4], [5]). Further, we find that these variables can also impact pitch accent type/alignment.

4. Conclusions

This analysis of three of the evidentiality/sourcehood conditions (contexts A, B, and D) suggests that speakers use intonational contours to signal degrees of evidence the speaker has about the facts, and, perhaps more importantly, about the hearer's beliefs and evidence about the facts. Preliminary evidence suggests that speakers use H-type accents when the speaker believes the hearer has no evidence (D), and L-type accents when the speaker believes the hearer has direct evidence (A). However, when the speaker and hearer transparently share the evidence (B), responses more mixed: speakers use L* predominantly (as in A), but some subjects choose H* instead.

A useful outcome of this experiment is the demonstration that the method used to prompt subjects was effective for eliciting consistent intonational contours. Although not every (non-consecutive) repetition for each subject was identical to other repetitions for that prompt, there was a high degree of within-speaker consistency, with only small amounts of variation, and a noticeable degree of agreement across speakers. Moreover, speakers' prosody changed depending on the condition and they did not simply repeat a favorite intonational contour for all prompts. This suggests that speakers understand the context and then produce what, for them, is the appropriate prosody for each of these conditions, with evidentiality/sourcehood as one of the conditioning factors.

Although this report involves only 10 subjects over three conditions, the results are promising for eliciting consistent results within speaker and condition. Moreover, since the stimulus manipulations were between contexts defined in terms of evidentiality and sourcehood, it finds evidence for a reliable connection between formal aspects of the semantic/pragmatic context on the one hand and specific intonational tones/contours on the other. More intuitive notions like sentence function (e.g., ”question” vs. ”statement”) are too coarse-grained to be useful in predicting the intonational contour; in particular, such notions could not be used to predict that context A and context B differ in the proportion of H* accents. (More broadly, there is no universal ”question” or ”statement” intonation; consider, e.g., the difference between WH-questions and Yes/No questions.)

However, one should be cautious in concluding that evidentiality/sourcehood are the primary factors at play in our results. For example, it should be noted that both A and B stimuli were punctuated with question marks, unlike those in D which were punctuated with periods. This may have influenced speakers to use certain contours in A/B differently from D. However, this does not nullify our conclusions; question marks do not always go with L* H-H% contours (e.g., WH questions and polar alternative questions), and additionally the intonational differences between A and B cannot be due to punctuation. To address this issue, future experiments will not include punctuation.

To further investigate this connection, we are in the process of increasing the size of the subject pool and labeling the remaining conditions (C and E). Additionally, we plan to investigate whether the naturalness of these contours depends on evidentiality/sourcehood in a perceptual experiment, using these cartoon prompts with audio recordings. Finally, further methods of describing the distinct contours will be investigated. For example, labellers had difficulty with labels that are less frequently encountered in laboratory speech (e.g., !H-L% vs. H-L%). Other methods for categorizing accents, such as the Tonal Center of Gravity, might shed more light on both alignment as well as scaling differences ([15]). Finally, we would like to highlight that this research will have an impact on further development of speech interface technologies. As automatic conversational agents become more widespread, users will begin to expect a nearly human experience. Without a clear understanding about what level of meaning/function is reflected by which set of prosodic categories, and which prosodic categories may map to a particular underlying meaning, developing algorithms to provide such an experience is impossible.

5. References

[1] J. Pierrehumbert and J. Hirschberg, “The meaning of intonational contours in the interpretation of discourse,” in *Intentions in Communication*, 1990.

[2] C. Gussenhoven, *On the Grammar and Semantics of Sentence Accent*. Dordrecht: Foris, 1983.

[3] D. Watson, M. Tanenhaus, and C. Gunlogson, “Interpreting pitch accents in online comprehension: H* vs. L+H*,” *Cognitive Science*, vol. 32, pp. 1232–1244, 2008.

[4] C. Gunlogson, “A question of commitment,” *Belgian Journal of Linguistics*, vol. 22, pp. 101–136, 2008.

[5] O. Northrup, “Grounds for commitment,” Ph.D. dissertation, UC Santa Cruz, 2014.

[6] J. Hirschberg, “Pragmatics and intonation,” in *The Handbook of Pragmatics*, 2004.

[7] C. Bartels, *The intonation of English statements and questions: A compositional interpretation*, 2014.

[8] E. Selkirk, *Phonology and syntax: the relationship between sound and structure*, 1984.

[9] K. Pruitt and F. Roelofsen, “The interpretation of prosody in disjunctive questions,” *Linguistic Inquiry*, vol. 44, no. 4, 2013.

[10] D. Farkas and F. Roelofsen, “Polarity particle responses as a window onto the interpretation of questions and assertions,” *Language*, vol. 91, no. 2, pp. 359–414, 2015.

[11] L. Winans, “Disjunction and alternatives in Egyptian Arabic,” 2015, ms., UCLA.

[12] E. O. Aboh, “Information structuring begins with the numeration,” *Iberia*, vol. 2, no. 1, pp. 12–42, 2010.

[13] G. Garding and A. Arvaniti, “Dialectal variation in the rising accents of American English,” *Laboratory Phonology 9*, 2004.

[14] M. E. Beckman, J. Hirschberg, and S. Shattuck-Hufnagel, “The original ToBI system and the evolution of the ToBI framework,” in *Prosodic Typology*, S.-A. Jun, Ed., 2005.

[15] J. Barnes, N. Veilleux, A. Brugos, and S. Shattuck-Hufnagel, “Tonal center of gravity: A global approach to tonal implementation in a level-based intonational phonology,” *LabPhon 3*, 2012.

"She has many… *cat*?": On-line Processing of L2 Morphophonology by Mandarin Learners of English

Valeria Peretokina[1], Catherine T. Best[1,2], Michael D. Tyler[1,3], Bruno Di Biase[2]

[1]MARCS Institute, Western Sydney University, Australia
[2]School of Humanities and Communication Arts, Western Sydney University, Australia
[3]School of Social Sciences and Psychology, Western Sydney University, Australia
{V.Peretokina, C.Best , M.Tyler, B.Dibiase}@westernsydney.edu.au

Abstract

The current study examined on-line processing of second-language (L2) morphophonology by Mandarin learners of English, as compared to native English listeners, in a cognitively demanding self-paced listening task. Participants' listening times (LTs) to target singular and plural nouns that varied in phonological complexity and grammaticality were measured and analysed. The phonological representation of the targets was revealed to influence L2 listeners' processing speed the most, while morphological plurality and grammaticality did not seem to modulate LT in native or L2 participants. Thus, listeners appeared to disregard grammatical violations in favour of utterance comprehension.

Index Terms: speech perception, self-paced listening, morphophonology, grammaticality, processing load.

1. Introduction

Certain second language (L2) morphological structures, such as grammatical plurality, are challenging to acquire. Adult L2 learners often demonstrate inconsistent use of these units in speech production [1, 2] and insensitivity to inflectional errors in on-line reading comprehension [3, 4]. For instance, Mandarin Chinese (ManC) learners of L2-English are reported to show variable L2 production of English plural morpheme <-s> [1] as well as unresponsiveness to the omission of plural inflections in real-time reading tasks [4]. Research on L2 acquisition has linked these findings to representational gaps in learners' native language (L1) morphology [2, 5] (e.g., lack of grammatical plurality in Mandarin [6]) or phonology [7] (e.g., phonotactic restriction against /s/ in coda position or consonant clusters in Mandarin [8]). Thus, according to the representational gap account, a morphophonological feature is assumed to be unattainable for native-like mastery if it does not have a counterpart in L1 grammar or if the inflection of an L2 word stem yields L1 phonotactic violations.

It has been found, however, that adult ManC learners of L2-English are able to show sensitivity to plural marker <-s> and to its omission in a phoneme monitoring task [9]. In that study, ManC intermediate learners of Australian English (AusE) and AusE monolinguals were presented with audio recordings of English utterances and instructed to detect a target phoneme /s/ (realised as [s]) in various morphological (singular vs. plural nouns) and phonological (singleton vs. cluster codas) contexts. In 'catch' utterances, /s/ indicating plurality was omitted. The findings revealed L2 listeners' sensitivity to violations of English morphophonological alternations, and native-like processing patterns of word-final /s/. Interestingly, the presence of a morpheme boundary in plural nouns was observed to quicken L2 perception of /s/, even though Mandarin lacks both morphological marking of grammatical plurality and syllable-final /s/ as singletons or in clusters. This indicates that L2 learners were able to use morphological information in comprehension similarly to native listeners, implying that difficulties with L2 morphophonology in speech production and on-line reading tasks may not inevitably result from gaps in L1 morphology.

The inconsistency in previous findings may stem from the range of experimental procedures used to test acquisition of L2 morphophonology, with the majority of past studies examining L2 speech production or reading comprehension and only investigating perception of spoken L2 using phoneme monitoring. Varying cognitive demands of the tasks may also contribute to the discrepancy in the observed results. Spontaneous speech production and self-paced reading tasks targeting on-line reading comprehension put participants under a high processing load by introducing time pressure and requiring an extensive working memory use. Phoneme monitoring, on the other hand, tests perception of a single feature and does not explicitly check stimulus comprehension, which reduces the amount of cognitive resources needed to accurately perform the task and allows participants to access their L2 morphophonological knowledge, even for structures that are not represented in the L1.

Given the lack of empirical evidence on perception of spoken L2 under high processing load, the present study seeks to test ManC listeners' sensitivity to L2 morphophonological variations in a self-paced listening task. Self-paced listening is comparable to phoneme monitoring due to the use of *spoken* rather than *written* stimuli, but is more cognitively demanding as, similarly to self-paced reading, it investigates L2 learners' real-time speech processing. To explore perception of L2 morphophonology depending on phonological representation, morphological complexity, and grammatical violations of the target words, singular nouns ending in [Vs] (e.g., *house*) and [Cs] (e.g., *chance*), plural nouns ending in [Cs] (e.g., *cats*), and ungrammatical plural nouns with omitted inflection (e.g., She has many *cat*) were included in the design. Listening time (LT) to each word group was measured, with longer LT signifying processing delays, and in case of ungrammatical plurals, reflecting sensitivity to morphological violations.

ManC participants are hypothesised to find singular nouns ending in [Vs] the easiest for processing, even under a heavy cognitive load. Even though this group of singular nouns deviates from Mandarin phonotactic rules by having /s/ in codas, it complies with Mandarin morphology and does not contain L1 phonotactically impermissible coda clusters. By contrast, singular nouns ending in [Cs] are presumed to induce longer LT due to a more complex phonological representation that is inconsistent with ManC phonotactic rules. As for L2 listeners' perception of plurals, competing predictions can be made. On the one hand, if morphological awareness of L2-

English grammatical plurality is not obscured by the cognitive demand of the task, ManC listeners are expected to exhibit a difference in the LT to plural nouns in comparison to both singular noun groups. As L2 listeners have previously demonstrated faster processing of plural than singular nouns in phoneme monitoring [9], similar LT patterns may occur in the present experiment. Alternatively, retrieval of knowledge about L2-English grammatical plurality, coupled with performing a presumably challenging on-line task, could be manifested by longer LT to plurals relative to both groups of singular nouns. If ManC participants are sensitive to grammaticality of the targets, L2 processing of the omission errors is hypothesised to result in the longest LT. Otherwise, if L2 listeners' ability to detect errors is hindered by the processing demands of the task, LT to ungrammatical plurals is expected to be comparable to the LT to correct plurals.

To manipulate the amount of cognitive load within the task, targets appeared in sentences, either finally or medially. A longer LT to targets in final relative to medial position is expected, as it will reflect the increase in processing demand associated with participants retaining preceding words of the utterance in their working memory.

2. Method

2.1. Participants

Forty-eight participants were recruited for the present study in Sydney, Australia. The test group comprised 24 ManC speakers with a length of residence (LoR) in Australia of less than one year (15 females, 9 males; M_{age} = 24 years, Range: 18–32 years; M_{LoR} = 5 months, Range: 1–10 months). The English proficiency level of ManC participants was categorised as intermediate, as determined by their IELTS scores. Twenty-four AusE monolingual speakers (15 females, 9 males; M_{age} = 25 years, Range: 18–39 years), who were born and raised in Australia and did not indicate proficiency in any languages other than English, acted as the control group. None of the participants had any hearing, speech or language impairments.

2.2. Materials

A list of 272 experimental and filler five-syllable stimulus sentences was created for this experiment (examples presented in Table 1). Experimental sentences contained target monosyllabic singular and plural nouns ending in /s/ that appeared utterance-medially or finally. Using the CELEX database [10], we selected only high-to-medium frequency words as targets to ensure that targets' semantic difficulty would have minimal effect on participants' performance in the task ($M_{\text{frequency}}$ = 93 per million, Range: 10–528 per million).

Each group of target nouns was divided into two predicted processing difficulty levels, 'easy' and 'difficult'. Difficulty level of the target words was determined separately for each noun category. Singular nouns were categorised to be 'easy' or 'difficult' depending on their phonological representation with the singular nouns ending in [Vs] classified as 'easy', while those ending in [Cs] identified as 'difficult'.

Plural marker <-s> always appeared in [Cs], given that when it is preceded by a vowel it is realised as the allophonic variant [z] (e.g., *keys*) rather than [s]. As /z/ does not exist in the ManC phonological inventory, the [z] allophone of <-s> would have presented additional difficulty for L2 processing. Thus, within the category of plural nouns the difficulty level distinction was based on grammaticality of the targets. Correct plurals were labelled as 'easy' in comparison to ungrammatical plural nouns with omitted inflections, predetermined as 'difficult'. The 'easy' targets were presumed to require shorter processing time than the 'difficult' targets.

Filler sentences either contained no target phoneme /s/ or /s/ appeared in any word/utterance position. Participants' responses to fillers were not included in statistical analyses.

A 26-year-old female monolingual AusE speaker read the sentences word by word in a neutral voice. The session was recorded in a soundproof booth at 44.1 kHz sampling rate using a Shure SM10A-CN headset microphone, a MOTU Ultralite-mk3 audio interface, and Cool Edit Pro 2.1 software.

Royalty-free photographs were found on-line for presentation with the audio stimuli. Images were processed to be approximately the same size, with the long side set to 400 pixels and preserved aspect ratio.

2.3. Design and procedure

Testing sessions were conducted individually in a sound-attenuated booth using an Acer TravelMate P653 laptop and Sennheiser HD 650 headphones connected to an Edirol UA-25EX external sound card. Each participant completed a background questionnaire and a computer-based self-paced listening task, programmed and presented in DMDX [11].

At the beginning of each trial participants were simultaneously presented with the first word of an utterance and a picture depicting the meaning of the sentence on a computer screen. The participants' task was to pace themselves through each utterance by pressing a space-bar on a computer keyboard, indicating that they understood a word and were ready to move on to the next one, until the end of an utterance. The picture was present on the screen throughout the duration of the utterance. Each trial was followed by a 'yes/no' comprehension question asking if the utterance and the picture matched. Mismatched trials constituted 20% of the total number of trials and were distributed evenly across different sentence types. Mismatches were introduced to ensure that participants were attending to comprehension when performing the task. Trial presentation order was randomised for each participant. Comprehension rate was calculated and LTs were analysed only for trials with correctly answered comprehension questions.

3. Results

3.1. Comprehension rate

Responses to picture-sentence matching questions were recorded and percentage of correct responses was calculated for each participant. Then, the AusE and ManC groups' mean comprehension rate scores were derived from these values.

Table 1. *Stimulus sentence types.*

Stimulus Sentences	**Morphological Structure**	**Difficulty Level**	**Utterance Position**	
			Medial	**Final**
Experimental	**Singular nouns**	**Easy:** ending in [Vs]	Her new *house* looked clean.	I would like more *ice*.
		Difficult: ending in [Cs]	Take a *chance* to learn.	Buy that candy *box*.
	Plural nouns	**Easy:** correct plurals	Fluffy *cats* looked cute.	I came for five *nights*.
		Difficult: omitted inflection	Many *book* lay there.	She has many *cat*.
Fillers			We had a picnic. The *sky* went dark grey.	

While all listeners showed a high comprehension rate (M_{AusE} = 90%, M_{ManC} = 85%), which was clearly above chance in matched and mismatched trials (M_{AusE} = 91%, M_{ManC} = 88%; M_{AusE} = 86%, M_{ManC} = 78%, respectively), an independent-samples *t*-test revealed a significant difference between the L1 and L2 groups, $t(382) = 4.07$, $p < .001$. Lower than native comprehension performance of ManC listeners was expected due to their intermediate L2-English proficiency. That is, we infer that the group difference in picture-sentence matching accuracy did not indicate L2 participants' lack of utterance comprehension and did not prevent us from obtaining a sufficient number of correctly perceived trials to conduct a reliable analysis of LT data.

3.2. Listening time (LT)

After discarding trials with incorrect responses to picture-sentence matching questions, the remaining LT means, shown in Figure 1, were submitted to a 2 (Language group: ManC, AusE) x 2 (Utterance position: medial, final) x 2 (Morphological structure: singular nouns, plural nouns) x 2 (Difficulty level: 'easy', i.e., singular nouns ending in [Vs], correct plural nouns; 'difficult', i.e., singular nouns ending in [Cs], plural nouns with omitted inflection) analysis of variance. Language group was a between-subject factor; utterance position, morphological structure, and difficulty level were within-subject factors. Analyses were conducted by subjects (F_1) and items (F_2).

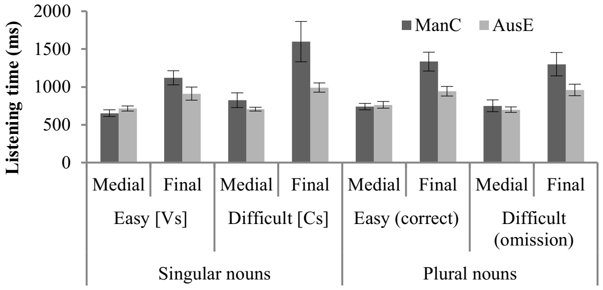


Figure 1: *Language group differences in listening time to target words as a function of utterance position, morphological structure, and difficulty level. Error bars represent standard error of the mean.*

As shown in Table 2, there were significant main effects of language group, with AusE participants exhibiting shorter LT (*M* = 837 ms) than ManC participants (*M* = 1041 ms), and utterance position, with longer LT to targets in final (*M* = 1147 ms) than medial position (*M* = 731 ms). A significant interaction between these two factors suggests that the utterance-final targets induced longer processing delays in L2 listeners than in L1 listeners (M_{ManC} = 1341 ms; M_{AusE} = 953 ms), while there were no significant group differences utterance-medially (M_{ManC} = 742 ms; M_{AusE} = 721 ms). A main effect of difficulty level reached significance in the subjects analysis, tentatively confirming our prediction that processing of 'difficult' targets would result in longer LT (*M* = 979 ms) than processing of 'easy' targets (*M* = 899 ms).

Two-way interactions between language group and difficulty level, and between morphological structure and difficulty level are also presented in Table 2. However, they will not be discussed in detail as the three-way interaction among language group, morphological structure, and difficulty level, shown in Figure 2, provides the most comprehensive account of the data. To examine the differences that led to the three-way interaction, simple effects tests with a Bonferroni correction were performed on the two simple two-way interactions between morphological structure and difficulty level for each language group. The data were collapsed across the utterance position factor, as it did not contribute to the interaction, and split by language group.

Table 2. *Significant main effects and interactions (in bold) in the analysis of variance.*

Main effects	
Language group	**F_1(1, 46) = 4.27, p = .044** **F_2(1, 136) = 3.21, p < .001**
Utterance position	**F_1(1, 46) = 49.4, p < .001** **F_2(1, 136) = 108.84, p < .001**
Difficulty level	**F_1(1, 46) = 4.45, p = .04** F_2(1, 136) = 2.53, p = .114
Interactions	
Language group * Utterance position	**F_1(1, 46) = 9.68, p = .003** **F_2(1, 136) = 29.59, p < .001**
Language group * Difficulty level	F_1(1, 46) = 3.85, p = .056 **F_2(1, 136) = 4.03, p = .047**
Morphological structure * Difficulty level	**F_1(1, 46) = 10.45, p = .002** **F_2(1, 136) = 7.41, p = .007**
Language group * Morphological structure * Difficulty level	**F_1(1, 46) = 5.01, p = .03** **F_2(1, 136) = 5.67, p = .019**

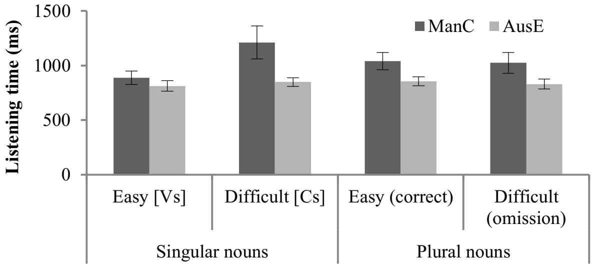


Figure 2: *Language group differences in listening time to 'easy' and 'difficult' morphological structures. Error bars represent standard error of the mean.*

For ManC participants (the dark grey bars in Figure 2), the simple two-way interaction between morphological structure and difficulty level was significant, $F_1(1, 23) = 8.37$, $p = .008$, $F_2(1, 136) = 7.54$, $p = .007$. Simple effects analyses revealed that the L2 group's LT to 'difficult' targets remained similar regardless of morphological structure, $F_1(1, 23) = 2.39$, $p = .136$, $F_2(1, 136) = 1.86$, $p = .175$, but was statistically different for 'easy' targets, $F_1(1, 23) = 25.63$, $p < .001$, $F_2(1, 136) = 5.76$, $p = .018$. That is, the LT to 'easy' singular nouns, i.e., singular nouns ending in [Vs] (*M* = 888 ms), was significantly shorter than the LT to 'easy' plural nouns, i.e., correct plurals (*M* = 1041 ms). Furthermore, a significant difference in L2 participants' LT was found when comparing the two difficulty levels within singular nouns, $F_1(1, 23) = 8.23$, $p = .009$, $F_2(1, 136) = 8.16$, $p = .005$. ManC listeners took more time to respond to 'difficult' singular nouns ending in [Cs] (*M* = 1212 ms) relative to 'easy' singular nouns ending in [Vs] (*M* = 888 ms). Hence, the shortest LT in ManC group was observed for singular nouns ending in [Vs], with this difference being confirmed both against singular nouns ending in [Cs], which constitutes the other level of the morphological structure factor for singular nouns, and against correct plurals, which represents the other pre-identified 'easy' target word type. As shown by the light grey bars in Figure 2, a simple two-way interaction between morphological structure and difficulty level for AusE participants was not significant, $F_1(1, 23) = 1.23$, $p = .14$, $F_2(1, 136) = .71$, $p = .4$. This indicates that there were no variations in L1 participants' LT depending on morphophonology and grammaticality of the targets.

The simple effects analyses allowed us to directly compare the ManC group's LT to correct plural nouns and singular nouns ending in [Vs] and revealed that plurals induced longer processing time in L2 listeners. However, it is unclear whether the increase in LT to plural nouns was observed due to their morphological complexity or their phonological representation (i.e., coda cluster). Therefore, a subsequent paired-samples *t*-test was carried out to compare ManC group's LT to plural and singular nouns ending in [Cs], as these two word groups differ only in morphological complexity. If it is indeed the grammatical inflection encoded by /s/ that makes plural nouns difficult for processing, we would expect ManC participants to demonstrate a longer LT to plural nouns than singular nouns ending in [Cs]. A lack of difference in LT to these two word groups, however, would suggest an L1 phonological rather than L1 morphological influence on L2 processing. No significant difference was uncovered by the *t*-test ($M_{\text{correct plurals}}$ = 1041 ms, $M_{\text{nouns [Cs]}}$ = 1212 ms), $t(47) = 1.38$, $p = .176$, which is consistent with the idea that it was the L1-violating phonological representation of the target words and not the L1-absent grammatical plurality that predominantly modulated the ManC group's processing speed.

4. Discussion

The aim of this study was to examine on-line processing of L2-English morphophonology by ManC listeners under a high cognitive load of a self-paced listening task. In particular, we investigated participants' LT to target singular and plural nouns depending on their phonological representation, morphological complexity, and grammaticality.

Our findings provide clear evidence for the cognitive load effect on perception of spoken L2. As anticipated, both participant groups experienced delays when processing utterance-final targets. Slower LT was presumably caused by the necessity to retain an entire utterance in working memory in preparation to answer a comprehension question, which in turn resulted in the increased cognitive load and hindered processing. The utterance-final delays were more prominent in the performance of the ManC than the AusE group. This observation highlights the difference between L1 and L2 speech processing, with the former being more automated and the latter requiring more cognitive effort.

Unlike the results of the previous phoneme monitoring study [9], which had demonstrated near-native performance of L2 learners, the present experiment revealed distinct differences between the processing patterns of the two participant groups. While AusE listeners did not show any LT variations depending on morphophonological complexity of the target words, ManC participants' processing was observed to be modulated by the phonological representation of the target nouns. In line with our hypothesis, L2 listeners demonstrated the shortest LT to singular nouns ending in [Vs], confirming L1 phonological and phonotactic influences on L2 speech perception. In comparison to singular nouns ending in [Vs], longer LTs were found for both singular and correct plural nouns ending in L1-impermissible [Cs], with no significant differences between the two. This finding suggests that it is indeed phonological representation that determined L2 processing speed, whereas grammatical plurality did not appear to contribute significantly to LT variability.

In accordance with previous studies on on-line reading comprehension [3, 4], L2 learners did not differentiate between grammatical and ungrammatical targets, which could potentially suggest a representational gap in their L2 grammatical system or difficulty accessing L2 morphological knowledge under time pressure. However, this interpretation is contradicted by the fact that, surprisingly, native listeners also did not demonstrate any changes in LT to plurals with omitted inflections in comparison to their correct counterparts. This unexpected finding could have been due to the experimental procedure itself, as participants appeared to concentrate on the comprehension of the utterances and responses to picture-sentence matching questions instead of attending to grammatical structures. Also, given that AusE participants come from an area of Sydney that shows high multi-cultural representation, it would seem likely that they have had substantial exposure to L2 learners of English of various backgrounds and proficiency levels prior to participating in the study. This could have made native listeners less sensitive to inflectional errors and potentially explain the lack of response to grammatical violations.

In sum, the present study provides an extensive account of L2 morphophonology perception in spoken language and verifies that the processing load of the task and the phonological representation of the targets have a pronounced effect on L2 learners' perceptual patterns. Contrary to the findings of the phoneme monitoring experiment [9], morphological structure and grammaticality did not influence L1 or L2 participants' LTs in the current study, indicating that processing of grammatical inflections and error identification are largely determined by the cognitive demands of the task. This also might suggest that listeners ignored grammatical errors and focused on attending to the meaning of the utterances. Our subsequent studies will test ManC participants with a longer LoR in Australia to explore whether L2 exposure improves morphophonology processing.

5. References

[1] G. Jia, and A. Fuse, "Acquisition of English grammatical morphology by native Mandarin-speaking children and adolescents: Age-related differences," *J. Speech, Lang. Hear. Res.*, vol. 50, no. 5, pp. 1280–1299, 2007.

[2] R. Hawkins and S. Liszka, "Locating the source of defective past tense marking in advanced L2 English speakers," in *The Interface between Syntax and Lexicon in Second Language Acquisition*, R. van Hout, H. Aafke, F. Kuiken, and R. Towell, Eds. Amsterdam: Benjamins, 2003, pp. 21–44.

[3] N. Jiang, E. Novokshanova, K. Masuda, and X. Wang, "Morphological congruency and the acquisition of L2 morphemes," *Lang. Learn.*, vol. 61, no. 3, pp. 940–967, 2011.

[4] N. Jiang, "Selective integration of linguistic knowledge in adult second language learning," *Lang. Learn.*, vol. 57, no. 1, pp. 1–33, 2007.

[5] I. M. Tsimpli and M. Dimitrakopoulou, "The interpretability hypothesis: Evidence from wh-interrogatives in second language acquisition," *Second Lang. Res.*, vol. 23, pp. 215–242, 2007.

[6] C. N. Li and S. A. Thompson, *Mandarin Chinese: A functional reference*, Oakland, CA: University of California Press, 1989.

[7] D. Lardière, "Attainment and acquirability in second language acquisition," *Second Lang. Res.*, vol. 22, pp. 239–242, 2006.

[8] S. Duanmu, *The phonology of Standard Chinese*, New York: Oxford University Press, 2007.

[9] V. Peretokina, C. T. Best, M. D. Tyler, J. A. Shaw, and B. Di Biase, "Perception of English codas in various phonological and morphological contexts by Mandarin learners of English," in *Proc. of the 18th Int. Congress of Phonetic Sciences, ICPhS 2015, 10–14 August 2015, Glasgow, Scotland* [Online], Available: https://www.internationalphoneticassociation.org/icphs-proceedings/ICPhS2015/Papers/ICPHS0603.pdf.

[10] R. H. Baayen, R. Piepenbrock, and H. Van Rijn, *The CELEX lexical database*, Philadelphia: University of Pennsylvania, 1993.

[11] K. I. Forster and J. C. Forster, "DMDX: A Windows display program with millisecond accuracy," *Behavior Research Methods, Instruments & Computers*, vol. 35, pp. 116–124, 2003.

Pause acceptability is predicted by morphological transparency in Wubuy

Brett J. Baker[1], *Rikke L. Bundgaard-Nielsen*[2]

[1]University of Melbourne, Australia
[2]MARCS Institute for Brain, Behaviour and Development,
Western Sydney University, Australia

bjbaker@unimelb.edu.au, rikkelou@gmail.com

Abstract

Research demonstrates that words in polysynthetic languages may be complex at the prosodic level. Little psycholinguistic research, however, has investigated the extent to which speakers of these languages are aware of word-internal structure, and whether morphological relations of different types affect the location of prosodic boundaries. We present an experiment testing the acceptability of words with embedded pauses, with speakers of the Australian language Wubuy. The results show that pauses are more acceptable at some word-internal morpheme boundaries than others. These boundaries are not consistently correlated with prosodic constituents, but are predictable on the basis of semantics and morphological productivity.

Index Terms: morphology, pause, polysynthesis, psycholinguistics

1. Introduction

Ever since the first descriptions of polysynthetic languages such as Aztec appeared in the 18th century, such languages have raised fundamental questions about the intersection of the *lexicon* (the word-stock of a language) and *syntax* (the rules for combining words into groups—phrases and sentences) as separate linguistic entities. For example, the 19th century linguist Müller, cited in [1] says, of 'incorporating' languages that they '*do away with the distinction between the word and the sentence*'. More recently, [2] argues that, for Cree and Dakota—if the complex word corresponds to any phonological unit at all—it is the phonological phrase. Similarly, [3], [4], and [5] argue that the complex word can constitute one or more intonation phrases in the Australian language Dalabon. These languages therefore call into question our understanding of the concept of 'word' in general, and highlight the need for current speech processing research to closely examine word processing (also) in speakers of polysynthetic languages.

Here, we test whether speakers of the Indigenous Australian language Wubuy (also known as 'Nunggubuyu' [6]) have access to information about the internal morphological structure of a range of complex 'words'. We do so by examining the perceptual acceptability of words with artificially-generated pauses in a range of word-internal positions. The results show that speakers accept 'words' containing internal pauses *only* if those pauses occur at what we call morphologically 'legal' positions—corresponding to the boundaries of what traditional linguistics would label morphemes, and in particular, morphemes which are consistently transparent in form and meaning.

Such a result suggests that Wubuy speakers have implicit knowledge of the internal structure of polysynthetic 'words' in Wubuy. To our knowledge, this study is the first to show this experimentally (though see [7] for a rare exception).

Our findings are consistent with previous research on naturally occurring pauses in the related language Dalabon [3] [4], as well as with preliminary acoustic analyses of Wubuy 'words' presented in [8] which demonstrated, like Dalabon, the existence of word-internal deliberate pauses associated with intonation phrase boundaries, c.f. also [14]. Together, these findings show that 'words' in these polysynthetic languages are crucially *not* like words in English, but prosodically like phrases: a collection of prosodic words, potentially separated by pauses and associated with one or more intonation phrases. This in turn suggests that speakers have knowledge of the internal structure of complex words, at the level of representation associated with prosodic structure. This kind of detailed knowledge of the internal structure of words is denied by many current theoretical models of morphology, such as [9].

2. Background

Wubuy is a language spoken in Eastern Arnhem Land in the Northern Territory, by perhaps 60 fluent L1 speakers with close affiliation to the remote settlement Numbulwar on the Gulf of Carpentaria. Wubuy is likely also an L2/L3 for a number of speakers in adjacent communities in north-east Arnhem Land, but the precise number of speakers (L1/L2/L3) is difficult to assess.

Wubuy is a polysynthetic language which allows both prefixing and suffixing. Wubuy words may be semantically very complex. In example (1), we see a verb inflected for two arguments, but also containing other elements that may be glossed with meanings appropriate to quantifiers and nouns (the first line shows the utterance as pronounced, the second line its underlying form):

(1) waraŋukulmuŋcalciraa
wu-wara-ŋu-kulmuŋ-jalcir-aa
NEUT-MULT-Ø-belly-be.hanging-PRES
'there are lots of fruits (lit. 'bellies') (there)'

As the translation indicates, words in Wubuy—like those in other polysynthetic languages—can have meanings usually associated with propositions in a language such as English. Indeed, in most of the world's languages, a propositional meaning like that in (1) *could not* be expressed by a single word; it would require minimally a phrase of some kind.

The evidence that strings such as (1) are 'words' comes primarily from two sources: phonology and distribution. Numerous phonological rules apply within strings like these [6], none apply outside of it. Secondly, while some sub-strings of (1) can occur independently, such as /kulmuŋ/ 'belly', many cannot. For instance, */wu-wara-ŋu-kulmuŋ/ cannot be a word, nor can any subpart of it apart from /kulmuŋ/.

‘Words’ in Wubuy can contain morphological relations of a range of types. Some morphological constituents, we hypothesise, are relatively transparent to speakers, for example, incorporated nouns such as *kulmuŋ* ‘belly’. This noun has the same form and meaning when used independently. It also has a consistent prosodic structure (it takes an initial stress accent). By contrast, the tense suffix *-aa* on the verb in (1) is not easily divisible from the verb stem itself, because verbs in Wubuy fall unpredictably into one of around a dozen of conjugation classes with distinct suffixal forms. Moreover, the meaning of this suffix, if it can be said to have one, is highly abstract: it serves to distinguish a tense category in the paradigm of this verb. We exploit these differences in morpheme transparency in the experimental design of the present study, where the stimuli (discussed below) exhibit a range of morphological relation types—some transparent, and others less transparent.

A recent study of polysynthetic ‘word’ production in the Australian language Dalabon, related to Wubuy, shows that speakers of this polysynthetic language pause deliberately—within words—at the boundaries of morphemes [3] [4]. In particular, pauses may follow inflectional prefixes, as long as prosodic requirements of minimum moraic size are met on both sides of the pause. An example is shown in (2), where ellipses indicate pauses of more than 150ms duration, and the hyphens represent morpheme boundaries:

(2) a. kaʔ-. . . ɻak-. . . mijan
he- wood- will.get
‘He will get firewood.’
b. ceʔ-. . . cark-. . . niŋijan
we.two- together- will.sit
'We will sit together.’

The significance of these examples for our understanding of polysynthetic ‘word’ structure and morphological theory more generally has not been systematically discussed, nor has speaker knowledge of ‘word’-internal structure been experimentally probed in such languages. In the following, we address this gap by presenting a perceptual acceptability experiment with speakers of Wubuy, testing the relative acceptability of ‘words’ with pauses inserted at a range of morphological boundaries (see the Method for a discussion of each of the different types).

3. Method

3.1. Experimental design

In order to test speakers’ knowledge of the internal structure of words, and in particular the differences in transparency discussed in 2, we conducted a two alternate forced-choice preference experiment consisting of 34 pairs of utterances involving complex (multi-stem) Wubuy words (Table 1).

3.1.1. Stimulus recording

The Wubuy stimuli were recorded by a highly literate, female L1 Wubuy speaker, aged in her late 50s. The speaker read each ‘word’ out loud from a computer monitor, until her production was fluent. She then repeated the ‘word’ at normal speaking rate five times, and either repetition 2, 3, or 4 was selected for inclusion in the perceptual experiment.

Acoustic analyses of utterances 2-3-4 of the recordings show that they contained silent pauses of approximately 250ms duration at a range of morphological boundaries—even in highly automatic, well-rehearsed, non-read speech. The duration of these pauses contrasts with the duration of stop constriction pauses in the same utterances, which were on average 50ms in duration [8].

In order to test the acceptability of (artificially extended) pauses at a range of morphological boundaries, a 500ms pause was inserted into a range of positions in each of the selected stimulus ‘words’ [8]. (The complete stimulus list, with translations, is shown in Table 1). We also retained the unmodified stimulus ‘words’ to allow for acceptability preference judgements to be given for words with no artificial pause inserted and words with pauses inserted at 'legal' and 'illegal' junctures.

3.1.2. Pause insertion

Pauses were inserted in four positions:

(A) between an incorporated noun and a verb stem;
(B) between two halves of a reduplicated verb-stem;
(C) between a bound verb root and finite root; or
(D) within a morpheme (such as *jina* ‘head’).

Type A is illustrated by example (1), as discussed above. Type B is illustrated in Table 1. While verbs reduplicate productively in Wubuy to indicate distribution in time, space or participants [6], there also exist many lexemes in the language which are ‘inherently’ reduplicated. These lexemes do not occur in an unreduplicated form, and do not have the meanings associated with productive reduplication. The verb /kucukuca-/ ‘tickle’, in Table 1, only ever occurs in this inherently reduplicated form. We therefore infer that the parts of this verb are not independently meaningful for speakers. Type C is illustrated by the verb form /wulṯa-/ ‘cut’; which in our stimulus list always occurs in the ‘hardened’ form /kulṯa-/ (because of a phonological rule preventing non-nasal sonorants from following non-continuants). This verb can be analysed historically as consisting of a stem /wul-/ and a finite verb root /ṯa-/, but synchronically neither of these elements is independently meaningful in Wubuy: it is a lexicalised compound. Unlike the Type A situation then, all three of Types B, C, D present speakers with word parts, separated by pause, which (we hypothesise) carry no independent meaning. The final relationship tested was that between the meaningless string ŋu-, also illustrated in (1), and a preceding or following stem. This string has no meaning in Wubuy: it is inserted by morpho-phonological rule and precedes stems beginning in underlying stops. We hypothesise that the occurrence of ŋu- on either side of a pause would have no effect on participant judgements of pause legality.

We inserted 500ms of silence to ensure that the artificially generated pause is on par with (or longer) than the majority of the pauses identified between lexical morphemes in the preliminary acoustic analysis [8]. This is also close to the average found across deliberate pauses in a recent study of French and German speakers [10]. We follow [8] in regarding the first type (A), as a 'legal' pause boundary, because the relationship between the parts of the word thus separated is transparent in the sense discussed in 2, and types (B, C, D) as 'illegal' pause boundaries, reflecting differences in semantic and morphological transparency of each of the parts. Note that when an utterance is so divided, neither part constitutes a licit word in Wubuy in this context. The utterance order was counter-balanced, and trial order pseudo-randomised.

Natural speech	Legal breaks	Illegal breaks
ŋa-ˌɻuluc-kulˈt̪aŋi 1sg-shade-cut.through.PC 'I cut the bough shade'	ŋa-ˌɻuluc-#kulˈt̪aŋi	ŋa-ˌɻuluc-kul#ˈt̪aŋi
a-ˌjina-ŋu-cuˈɻang 1sg/2sg-head-Ø-push.NFUT 'I'll push your head'	a-ˌjina-#ŋu-cuˈɻang a-ˌjina-ŋu-#cuˈɻang	
ŋan-ˌcina-kaˈɭaɭic 1sgIRR-head-wet 'my head will get wet'	ŋan-ˌcina-#kaˈɭaɭic ŋan-#ˌcina-kaˈɭaɭic	ŋan-ˌci#na-kaˈɭaɭic
ŋani-ˌjina-ŋu-ˌkucukuˈcaani 3MASC/1sg-head-Ø-tickle.PC 'He tickled my head'	ŋani-ˌjina-#ŋu-ˌkucukuˈcaani ŋani-ˌjina-ŋu-#ˌkucukucaani ŋani-#ˌjina-ŋu-ˌkucukucaani	ŋani-ˌjina-ŋu-ˌkucu#kuˈcaani

Table 1: *Stimulus list in Wubuy with translations. '#' indicates location of inserted pause of 500 ms.*

3.2. Participants

We recruited 14 L1 speakers of Wubuy (one male; one participant had grown up using Wubuy but was now dominant in Kriol). The participants ranged in age from approximately 40 to approximately 65. All spoke L2/L3 Kriol and English to varying levels of competence. We excluded four participants because they failed to understand the task.

The participants were informed that they would hear pairs of utterances from a familiar Wubuy speaker. Participants heard each pair of utterances through headphones from a laptop. For each pair of utterances, the listeners were instructed to choose the one which sounded 'best' to them, by means of a hand gesture or by saying 'first one' or 'last one'. They were allowed to listen to each pair as many times as they liked before making their decision. Testing took place in quiet homes in either Darwin or Numbulwar. All participants were compensated for their time and effort by a payment of $50.

3.3. Predictions

We predicted that natural, fluent utterances ('N' in what follows) and utterances with pause at a legal boundary ('L') would be preferred over utterances with pause at an illegal boundary, irrespective of the type ('IL'). We also predicted that natural utterances would be predicted over utterances with a pause inserted. However, the literature [11] suggests that speakers might actually *prefer* complex utterances to have pauses, to aid in processing.

4. Results

Mean results for the 10 participants are presented in Figure 1. In order to determine whether the preference patterns differed from chance performance, we conducted three one-sample *t*-tests against chance (50%). The results are consistent with our first two predictions: Natural (N) utterances are preferred over illegal (IL) pause-inserted utterances (79%; $p < .001$). Legal (L) pause-inserted utterances are also preferred over illegal (IL) pause-inserted utterances (85%; $p < .001$). The preference pattern for natural utterances (N) over legally modified utterances (L) did not differ from chance (61%; $p = .068$).

To test our prediction that natural (N) and legal (L) utterances would be preferred over illegal (IL) utterances, conducted a one-way ANOVA. Test assumptions of normality and homogeneity of variance were satisfactory, and the results indicate that the participants did indeed respond differentially to the N > IL, L > IL, and N > L conditions ($F(2, 27) = 6.352$, $p = .005$). Bonferroni post-hoc comparisons revealed that the preference pattern for N > L differed from the preference pattern for both N > IL ($p = .044$) and L > IL ($p = .006$). The preference pattern for N > IL and L > IL did not differ ($p = 1.000$), indicating that the strength of the preference for N and L utterances (over IL) was comparable.

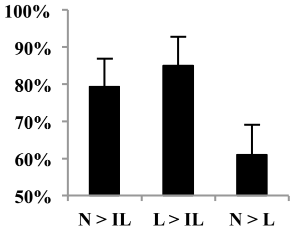


Figure 1. *Mean 'Word' preferences. N= natural utterance; L = pause inserted at a 'legal' boundary; IL = pause inserted at an illegal boundary. Error bars indicate positive SD.*

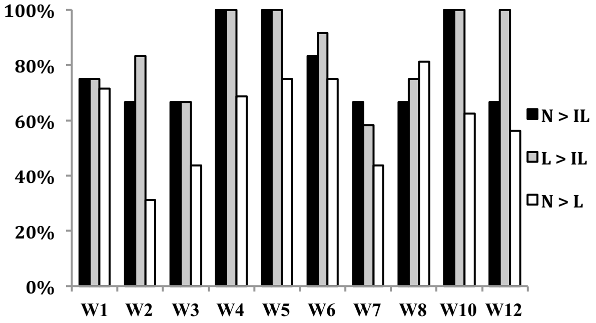


Figure 2. *Individual 'word' preferences for the 10 participants. W = participant ID. N = natural utterance; L = pause inserted at 'legal' boundary; IL = pause inserted at illegal boundary.*

Figure 2, above, presents the individual preference results. As can be seen from this figure, the individual preference for the L > N utterances differed: seven participants preferred the natural utterances (N > L), while three listeners preferred the legal utterances (i.e. with artificially inserted pause) over the natural ones. This is consistent with our speculation (above) that speakers might in fact prefer complex words to have internal pauses, because they are easier to process, as suggested, for English utterances, by [11].

5. Conclusions

The current study shows that Wubuy speakers make consistent judgements about the locations of deliberate pauses within complex words. The results also show that word-internal pauses are in fact *preferred* by some speakers, over utterances lacking internal pause. Preliminary data reported in [8] suggest that this reflects the fact that Wubuy speakers often pause within words in speech. However, pauses are not acceptable in all potential positions within words: pauses are only acceptable in instances where the pause is between what a traditional analysis would regard as morphemes (i.e. consistent associations of phonological material with some meaning). Moreover, not all morpheme boundaries are equally acceptable: only in those cases where each component thus separated is meaningful, are inserted pauses acceptable at such a boundary. Our explanation for the consistent non-acceptance by speakers of forms such as /ŋani-jina-ŋu-kucu#kucaani/ and /ŋa-ɻuluc-kul#t̪aŋi/ is that the pause separates a word into strings that have no meaningful analysis.

One question which presents itself is whether participants are simply reacting to the location of *prosodic* (rather than morphological) boundaries within words. That is, is it the case that pauses are acceptable at the boundaries of prosodic words, regardless of the locations of such constituents with respect to morphology? Teasing apart the contribution of prosody and morphology is not straightforward, because, as in the vast majority of Australian languages for which we have detailed descriptions, morphological structure crucially determines the locations of metrical constituents such as feet and prosodic words [12]. Moreover, the prosodic description of Wubuy is in its infancy. Our impressionistic observations of the locations of stressed syllables are given in Table 1. Based on this, and the existing analyses of related languages such as Dalabon [5] and Ngalakgan [13], we infer that Wubuy has trochaic (left-headed) metrical feet aligned preferentially with the initial syllable of lexical stems (nouns and verbs) but also with the penultimate syllable (c.f. [14]). In verbs, the penultimate syllable tends to be the one associated with primary stress (that is, with the pitch accent associated with an Intonational Phrase: [5]). This is true of most of our data, with the exception of the verb /cuɻaŋ/, which we hear with final stress (for reasons that are obscure at this point). In any case, note that there is no regular correlation between the acceptability of pause and the locations of primary or secondary stress accents. In other words, the acceptability of pause cannot be predicted from the prosodic structure of the word alone.

These results demonstrate that speakers of Wubuy, and by extension speakers of other, similar languages, have access to the internal structure of complex words, not just the locations of prosodic constituents, but also the ways in which these constituents are associated with the semantics of the whole word. This may seem unsurprising, but it is unexpected from the point of view of many current models of morphology, which deny the cognitive reality (or analytical usefulness) of the traditional notion 'morpheme', e.g. [9].

Our findings thus complement existing work on the prosodic structure of polysynthetic Australian languages such as Dalabon [5], showing that words can consist of multiple instances of Prosodic Word (each associated with a lexical stem such as noun or verb), and thus constitute intonational constituents of type 'IP' (intonational phrase) or 'AP' (accentual phrase). Moreover, for Dalabon, [5] argues that the presence of pause and internal pitch movements are diagnostic of *word-internal* AP or IP boundaries in some cases. As in Dalabon, Wubuy speakers also insert deliberate pauses mid-word in elicited productions [8], indicating that such words are likewise complex at the level of Intonation Phrase.

To conclude, we submit that the behaviour of Wubuy speakers in this experiment, taken together with existing prosodic descriptions, casts serious doubt on the proposition that constructs like those in (1) are 'words' in the traditional sense: prosodically free, lexically idiosyncratic items with a rigid internal structure not necessarily accessible to speakers' consciousness. By contrast, 'words' in Wubuy can contain a number of items which are prosodically free in the sense that they can be followed or preceded by deliberate pauses, have a semantic structure which is compositional rather than idiosyncratic, and have a structure which is accessible to speaker consciousness to the extent that speakers are able to judge whether pauses are acceptable or not depending on their location with respect to this structure. The implication here is that our views of what constitutes a 'word' are in need of some revision; c.f. [15]. The only sense in which constructs such as (1) are word-like is possibly the distribution of phonological rules: Wubuy has numerous phonological rules which apply between morphemes within a word, but none that apply across word boundaries. Therefore, it is only in the sense 'phonological domain' that Wubuy complex words satisfy the criteria for word-hood, c.f. [2]; in other respects, they are phrasal.

6. Acknowledgements

We gratefully acknowledge our Wubuy participants. We also thank Elise Bell for her assistance. This study was funded by the ARC (DP130102624).

7. References

[1] Brinton, D. G., "On polysynthesis and incorporation as characteristics of American languages," Proc. Am. Phil. Soc., 23(121):48-86, 1886.

[2] Russell, K., "The 'word' in two polysynthetic languages," in T. A. Hall, & U. Kleinhenz (Eds), Studies on the Phonological Word, Benjamins, 203-221, 1999.

[3] Fletcher, J., Evans, N., & Ross, B., "Pausing strategies and prosodic boundaries in Dalabon," Proc. SST, 436-9, 2004.

[4] Evans, N., Fletcher, J., Ross, B., "Big words, small phrases," Linguistics, 46:89-129, 2008.

[5] Fletcher, J., "Intonation and prosody in Dalabon," in S.-A. Jun (Ed) Prosodic Typology II, 257-272, OUP, 2014.

[6] Heath, J., Functional Grammar of Nunggubuyu, AIAS, 1984.

[7] Rice, S., Libben, G., and Derwing, B., "Morphological representation in an endangered, polysynthetic language" Brain & Lang. 81(1): 473-486, 2002.

[8] Baker, B & Bundgaard-Nielsen, R.L., "Polysynthetic words are like sentences: evidence from pause placement and acceptability," Talk presented to ALS, Sydney, 2015.

[9] Stump, G. T., Inflectional Morphology: A Theory of Paradigm Structure. CUP, 2001.

[10] Trouvain, J., Fauth, C., & Möbius, B., "Breath and non-breath pauses in fluent and disfluent phases of German and French L1 and L2 read speech," Proc. Speech Prosody, 31-35, 2016.

[11] MacGregor et al., "Listening to the sound of silence: Disfluent silent pauses in speech have consequences for listeners," Neuropsychologia, 48:3982-3992, 2010.

[12] Baker, B., "Word structure in Australian languages," In H. Koch and R. Nordlinger (Eds), World of Linguistics: Australia, Mouton, 137-211, 2014.

[13] Baker, B., Word structure in Ngalakgan, CSLI, 2008.

[14] Hore, M., "Syllable length and stress in Nunggubuyu," In B. Waters (ed.) Australian Phonologies: Collected Papers 5, 1-62, 1981.

[15] Haspelmath, M., "The indeterminacy of word segmentation and the nature of morphology and syntax," Folia Linguistica 45:31-80, 2011.

Morphological status and acoustic realization: Findings from New Zealand English

Julia Zimmermann

Heinrich-Heine-University Düsseldorf

julia.homann@uni-duesseldorf.de

Abstract

This paper investigates the acoustic realization of morphemic and non-morphemic S in New Zealand English. A corpus study is reported that examines the role of morphological structure in fricative duration. Multiple linear regression is used to isolate these effects, which are then compared to previous findings on the homophony of morphemic and non-morphemic S in General American English. The results demonstrate the importance of morphological structure in speech production.

Index Terms: phonetic detail; morphological structure; English; homophony

1. Introduction

Recent research on lexeme homophony has shown that seemingly homophonous lexemes actually differ in phonetic details such as duration and vowel quality (e.g. [1], [2]). This poses a challenge to traditional models of speech production which locate frequency information at the level of the phonological form, and which postulate that phonetic processing and the module called 'articulator' do not have access to any information regarding the lexical origin of a sound (e.g. [3], [4]). Leaving stylistic and accentual differences aside, a certain string of phonemes in a given context should therefore always be articulated in the same way according to these models, irrespective of its morphemic status, and only show phonetic variation originating from purely phonetic sources such as speech rate or context.

The findings on lexemes prompt the question of whether similar differences also hold for allegedly homophonous affixes (instead of free lexemes). Early experimental research found some evidence that morphemic and non-morphemic sounds may differ acoustically. Walsh & Parker [5] carried out a production experiment and measured the duration of /s/ in three pairs of monomorphemic words and their homophonous counterparts that contained a final morphemic /s/ (e.g. *lapse* versus *laps*). In two out of three experimental conditions they found a small difference in the means of the two different kinds of /s/, with morphemic, i.e. plural /s/, being on average nine milliseconds longer than non-morphemic. Similarly, Losiewicz [6] investigated the acoustic difference between morphemic, i.e. past tense, /d/ and /t/, and non-morphemic /d/ and /t/ using an experimental setup, and also found durational differences between the two sets of sounds, with past tense /d/ and /t/ being longer than non-morphemic /d/ and /t/. Both of these studies, however, only considered very small data sets and did not control for all potentially confounding covariates that might have influenced the duration of the segments.

More recently, Plag, Homann & Kunter [7] conducted a corpus study to investigate the duration of S (that is [s] or [z]) as non-morphemic instances and as markers of plural, genitive, genitive plural, 3rd person singular and the cliticized forms of *has* and *is* in General American English. They used multiple regression modelling to control for pertinent covariates and found systematic differences in duration between the different kinds of S. However, their results went in the opposite direction of those of Walsh & Parker [5], with non-morphemic S being longer than the morphemic S. Furthermore, within the group of morphemic S, the affixes were found to be systematically longer than the clitics.

Seyfarth et al. [8], like Walsh & Parker [5], find morphemic S to be longer than non-morphemic S when considering homophonous word pairs such as *lacks* and *lax*. They used an experimental setup in which pairs of participants read out naturalistic dialogues that served as carriers for the words under investigation. The effect they find is ascribed to phonetic paradigm uniformity, where "[a] word's phonetic realization is influenced by the articulatory plans of its morphological relatives" [8], i.e. the articulatory plan of the base of the complex word (*lack*) affects the phonetic realization of the complex word (*lacks*), while no such effect is available for the simplex word (*lax*). The authors attribute the differences between their own and Plag, Homann & Kunter's findings to the sample size of the corpus study, which they suspect to have caused an imbalance in terms of the syntactic positions in which the items occur.

These divergent findings and open questions call for further evidence about the nature of durational differences between morphemic and non-morphemic S in English.

2. Morphemic and non-morphemic S in New Zealand English

The present study extends the research on the acoustic properties of affixes by looking at the behavior of S in a different variety of English, namely New Zealand English. Using over 6,900 items from the Quakebox corpus [9], the duration of morphemic and non-morphemic S is investigated in order to test whether New Zealand English shows the same systematic durational differences as found for General American English by [5], [7] or [8], whether it displays a different pattern or whether it displays no difference at all.

If there are indeed the same differences in the durations of the different S to be found as in Plag, Homann & Kunter [7], this would underpin the notion that the acoustic realization of English S is influenced by its morphemic status and furthermore strengthen the corpus-based findings in [7].

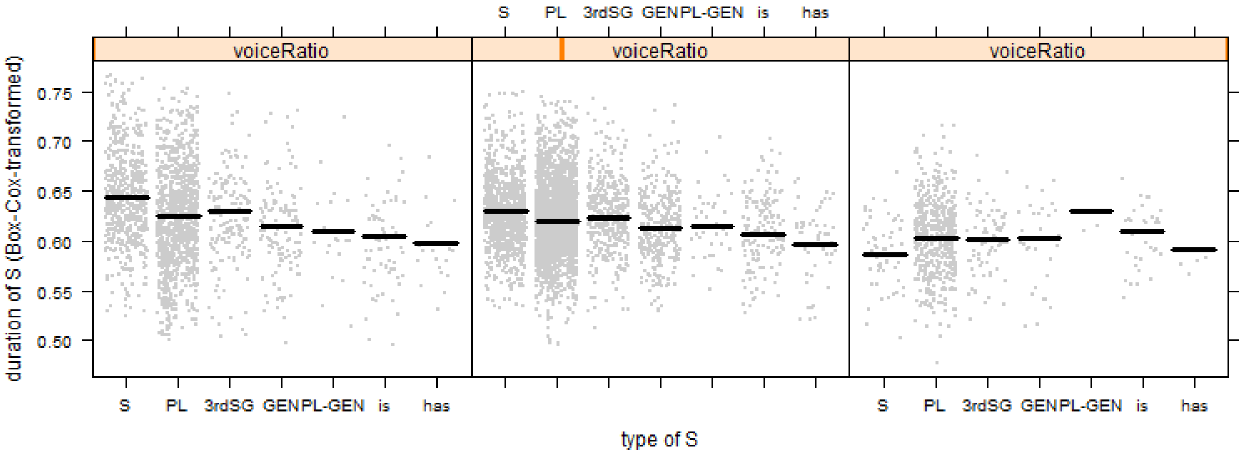


Figure 1: *Interaction plot for estimated box-cox transformed durations of S by amount of voicing and type of S. Model estimates are represented by solid horizontal lines, distribution of actual measured values is represented by grey dots. Panels represent voiceless (0-12% voicing), partially voiced (12-62% voicing) and fully voiced (62-100% voicing) items from left to right.*

2.1. Data

Following Plag, Homann & Kunter [7], non-morphemic instances of word-final S and S as a marker of plural, genitive, genitive plural, 3rd person singular and the cliticized forms of *has* and *is* were included in the analysis. Examples of each type of S in context are given in (1).

(1) non-morphemic: a *series* of aftershocks
(2) plural: there were huge *clouds* of dust
(3) genitive: my *family's* houses were okay
(4) genitive plural: we went to my *parents'* house
(5) 3rd person singular: something *falls* on it
(6) *has*-clitic: all this *mud's* come up
(7) *is*-clitic: the *lift's* broken

Items, i.e. morphemic S and the base it is attached to, or non-morphemic S and the word of which it forms the final segment, were sampled from the Quakebox Corpus [9]. This corpus is a collection of transcribed audio and video recordings of Cantabrians talking about their experiences in two major earthquakes that occurred in Christchurch in 2010 and 2011. The interviews were recorded in the 'QuakeBox', a shipping container which had been converted for use as a transportable recording studio. It was placed in different locations across the city of Christchurch. Audio was sampled at 48kHz stereo, using two Earthworks SR30 microphones (one headset microphone worn by the participant, one ceiling microphone inside the booth) and a USBPre2 microphone amplifier on a laptop computer running Audacity. At the time of data collection for the current study, 85 hours of high-quality recordings containing over 830,000 word tokens produced by 774 speakers were available. To keep the dataset free from potential dialectal differences, only those speakers who had identified themselves as native speakers of New Zealand English with a European background were included in this study (N=368). Using the corpus' automatically aligned phonetic transcriptions, all S-final words that were not followed by an S-initial word were extracted from the relevant recordings. Irregular forms, grammatical categories except for indefinite pronouns, brand and place names and items ending in in [ɪz] or [əz] were excluded manually. Due to the nature of Quakebox, the initial dataset of about 15,000 items was highly imbalanced in terms of type frequencies, with e.g. *house* contributing more than 2,000 items. For balancing purposes, only up to 25 randomly selected tokens were considered per type, leading to a reduced dataset of about 7,600 items. With the help of a Praat [1] script, relevant acoustic measures such as duration and voicing were extracted automatically.

In order to validate the automatic segmentations, the segmentations of 240 randomly selected items were checked manually using Praat [10]. Several different measurements for frequential center of gravity of the S in these items were considered as indicators for the reliability of the automatic segmentations, as any non-fricative material included in the S would have an effect on its frequential center of gravity. For each type of frequential measurement, the respective values for the automatic and the manual segmentations were then plotted against each other. Visual inspection of these plots showed clear tendencies where the automatic segmentations deviated most from the manual segmentations. For instance, frequential center of gravity weighted by the absolute spectrum based on the automatic segmentations yielded values from 1,000 to 13,500 Hz, while using the respective manual segmentations yielded values from 3,000 to 10,000 Hz. Therefore, any items with a frequential center of gravity weighted by the absolute spectrum ranging below 3,000 or above 10,000 Hz were excluded from the reduced dataset. The final set contained 7,081 tokens (from 1,879 types) for which the automatic segmentations were considered reliable.

2.2. Results

Linear mixed effects regression with a number of pertinent covariates (such as frequency, speaking rate, phonetic environment, etc.) was used to predict the duration of the S. The distribution of the durations of the S was slightly skewed and thus lacked linearity. This could have yielded unreliable estimates in linear regression, since one of the central assumptions of any linear regression model is a linear relationship between the dependent and independent variables. To alleviate this problem, the durations were Box-Cox transformed ([11], $\lambda = 0.2222222$).

Table 1: *Significance levels of individual contrasts between voiceless types of S as found in this study (Significance codes: '***' p < 0.001, '**' p < 0.01, '*' p < 0.05, '.' p < 0.1). Durations range from longest in leftmost column/top row to shortest in rightmost column/bottom row.*

	S	PL	3SG	GEN	PL-G	is	has
S	///	***	***	***	***	***	***
PL		///		**	.	***	***
3SG			///		*	***	***
GEN				///		*	*
PL-G					///		
is						///	
has							///

Table 2: *Significance levels of individual contrasts between voiceless types of S as found by Plag, Homann & Kunter [7] (Significance codes and duration range identical to Table 1).*

	S	PL	3SG	GEN	PL-G	is	has
S	///	**	*	***	**	***	***
PL		///				*	*
3SG			///			*	*
GEN				///			
PL-G					///		
is						///	
has							///

Models were fitted starting out with a fully specified model that contained all predictors that could be expected to have an effect on the (transformed) duration of S according to previous research (e.g. [12], [13], [14]). Stepwise exclusion of insignificant predictors, following the same simplification procedure as employed by Plag, Homann & Kunter [7], led to the final model, which showed a significant random effect of Speaker and significant main effects of Speaking Rate, Number Of Consonants in the rhyme of the final syllable of the item, Number Of Syllables in item, Duration Of The Base, Type Of Following Segment (pause, affricate, approximant, fricative, nasal, plosive, vowel), Duration Of Following Sound, Duration Of Preceding Sound, number of uses of item in Previous 30s of speech, log of Item Frequency in Quakebox, log of frequential Center Of Gravity by absolute spectrum and an interaction between Type Of S and Amount Of Voicing Of S (i.e., the ratio of voiced frames to total number of frames in the S). All effects go in the expected directions.

Figure 1 displays the average model estimates for the interaction between Voicing and Type Of S. The panels represent voiceless (left panel, 0-12% voicing), partially voiced (middle panel, 12-62% voicing) and fully voiced (right panel, 62-100% voicing) items, while type of S and transformed duration of S can be found on the x- and y-axis, respectively. As can be seen in the two leftmost panels, voiceless and partially voiced non-morphemic S are longer than most other types of S in those two conditions, while suffix S tend to be longer than clitic S. The back-transformed estimated mean durations for the group of voiceless S range from 98ms (*has*) to 139ms (non-morphemic S). The significance levels of all individual contrasts between voiceless types of S can be found in Table 1. The three ranges for the amount of voicing used in Figure 1 are based on the distribution of voicing in the dataset, which displays three main peaks, as illustrated in Figure 2.

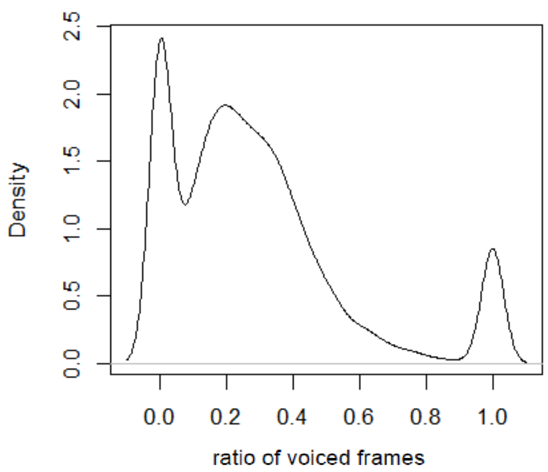


Figure 2: *Density distribution of amount of voicing in S*

3. Discussion

The study presented in this paper provides evidence for the existence of correlates of morphological structure in the acoustic signal. The duration of S in New Zealand English is dependent on morphological status. These findings clearly pattern with those by Plag, Homann & Kunter [7] for American English. In fact, the contrasts between the different kinds of voiceless morphemic S are even more pronounced in New Zealand English than they are in American English (cf. Table 2), with four additional significant contrasts in the former compared to the latter. The estimated durational difference between non-morphemic S and *has*-clitic S in this study (41ms) is also very close to the one observed by Plag, Homann & Kunter (38ms). Altogether, this study was able to replicate their results using a dataset more than ten times the size of the dataset of the original study. In these dimensions, potential imbalances in terms of syntactic position of the items, as suspected by Seyfarth et al. [8] about Plag, Homann & Kunter's dataset, should not be an issue.

At a very general level, these findings can be interpreted as support for the idea that there is morphological information in the phonetic signal, i.e. in postlexical stages of speech production. This goes against the assumptions of standard feed-forward formal theories of morphology–phonology interaction (e.g. [12], [16]). In these models, allomorphy is determined at a particular phonological cycle inside the lexicon, and at the level of underlying representations. Once the correct underlying form is derived, the morphological boundary of the respective cycle is erased (a process called 'bracket erasure', see [12], [16]) and the form leaves the lexicon. All further phonological processes are relegated to another module called 'postlexical phonology' and later to the articulatory component, neither of which have access to morphological information. According to my findings, it is possible to trace information about the structural status of a sound in the acoustic signal. Thus, the observed differences between the different Types Of S call into question the distinction between lexical and post-lexical phonology [16], which in turn would have important implications for theoretical mechanisms like bracket erasure and cyclic application of morpho-phonological rules.

At the theoretical level, these findings further challenge standard assumptions in models of speech production. Well-established models of speech production and the mental lexicon seem unable to accommodate my findings. Levelt, Roelofs & Meyer [17], for example, assume that pre-programmed gestures, which are stored in a syllabary, are executed by the articulator for the discrete syllables and segments of a language, which are phonologically represented. However, the articulator cannot provide a pre-programmed gesture for each syllable of a language if different meanings cause differences in these gestures. It is problematic that in such models, morphologically dependent sub-phonemic detail is not part of these representations. Such detail would need to be accounted for by purely phonetic factors that influence articulatory implementation such as speech rate [3]. For the duration of S, such an account is ruled out, as the effect of the type of S persists besides purely phonetic influences.

To summarize, both phonological theory and extant psycholinguistic models fail to provide a convincing explanation for the existence of morphological structure in the acoustic signal that I find in my data.

4. Conclusion

This paper has systematically investigated the relationship between morphemic status and phonetic implementation of homophonous affixes and their non-morphemic counterparts. This was done using natural conversation data. The analysis has yielded important evidence on the question of affix homonymy, revealing that phonologically homophonous bound morphemes can be phonetically distinct, and that morphemic and non-morphemic S may differ as well. This is unpredicted by current linguistic and psycholinguistic theories of lexicon and grammar. Further studies are certainly called for to be able to develop new models of the mental lexicon and of the relationships between morphology, phonology and phonetic implementation.

Furthermore, additional research is needed to address the many questions the present study raises. If there are indeed systematic differences between the different types of S in speech production, one would also like to know whether language users are influenced by these differences in perception. The difference in mean estimated duration between plural and *has*-clitic S amounts to 41ms. This difference lies well above the threshold for differentiating two fricative sounds that only differ in duration (e.g. [18], [19]) and translates to the average plural S being more than 1.4 times as long as the average *has*-clitic S.

5. Acknowledgements

This research was funded by the Deutsche Forschungsgemeinschaft (Forschergruppe 2373 'Spoken Morphology', TP3 VAR, PL 151/8-1 and ZP PL 151/7-1), which I gratefully acknowledge.

I would also like to thank Jen Hay and everyone at UC's NZILBB in Christchurch for their support in carrying out this study.

6. References

[1] Drager, K. Sociophonetic variation and the lemma. Journal of Phonetics 39(4):694–707. 2011.

[2] Gahl, S. *Time* and *thyme* are not homophones: The effect of lemma frequency on word durations in spontaneous speech. Language 84(3):474–496. 2008.

[3] Levelt, W. J. M. Speaking: From intention to articulation. Cambridge, MA: The MIT Press. 1989.

[4] Levelt, W. J. M. and Wheeldon, L. R. Do speakers have access to a mental syllabary? Cognition 50(1):239–269. 1994.

[5] Walsh, T. and Parker, F. The duration of morphemic and non-morphemic /s/ in English. Journal of Phonetics 11(2):201–206. 1983.

[6] Losiewicz, B. L. The effect of frequency on linguistic morphology. Ph.D. dissertation, University of Texas, Austin. 1992.

[7] Plag, I., Homann J. and Kunter, G. Homophony and morphology: The acoustics of word-final S in English. Journal of Linguistics, Available on CJO 2015 doi:10.1017/S0022226715000183. 2015.

[8] Seyfarth, S., Garellek, M., Malouf, R., and Ackerman, F. Acoustic differences in morphologically-distinct homophones. Talk presented at the 3rd American International Morphology Meeting. Amherst, MA. 2015.

[9] Walsh, L., Hay, J., Bent, D., Grant, L., King, J., Millar, P., Papp, V. and Watson, K. The UC QuakeBox Project: creation of a community-focused research archive. New Zealand English Journal, 27:20–32. 2013.

[10] Boersma, P. and Weenink, D. J. M. Praat: Doing phonetics by computer (Version 4.3.14) http://www.praat.org. 2013.

[11] Box, G. E. P. and Cox, D. R.. An analysis of transformations. Journal of the Royal Statistical Society, Series B 26(2):211–252. 1964.

[12] Hanique, I., Ernestus, M. and Schuppler, B. Informal speech processes can be categorical in nature, even if they affect many different words. The Journal of the Acoustical Society of America 133(3):1644–1655. 2013.

[13] Pluymaekers, M., Ernestus, M. and Baayen, R. H. Lexical frequency and acoustic reduction in spoken Dutch. Journal of the Acoustical Society of America 118(4):2561–2569. 2005.

[14] Pluymaekers, M., Ernestus, M., Baayen, R. H. and Booij, G. Morphological effects in fine phonetic detail: The case of Dutch -igheid. In Cécile Fougeron (ed.), Laboratory phonology 10 (Phonology and Phonetics), 511–531. Berlin and New York: Mouton de Gruyter. 2010.

[15] Chomsky, N. and Halle, M. The sound pattern of English. New York: Harper and Row. 1968.

[16] Kiparsky, P. Lexical morphology and phonology. In In-Seok Yang (ed.), Linguistics in the morning calm: Selected papers from SICOL, 3–91. Seoul: Hanshin. 1982.

[17] Levelt, W. J. M., Roelofs, A. and Meyer, A. S. A theory of lexical access in speech production. Behavioral and Brain Sciences 22(01):1–38. 1999.

[18] Klatt, D. H. Linguistic uses of segmental duration in English: Acoustic and perceptual evidence. Journal of the Acoustical Society of America 59(5):1208–1221. 1976.

[19] Shatzman, K.B., & McQueen, J.M. Segment duration as a cue to word boundaries in spoken-word recognition. Perception & Psychophysics, 68(1):1-16. 2006.

A Noise-Robust Linear Prediction Analysis for Efficient Speech Coding

Aadel Alatwi[1], *Stephen So*[1], *Kuldip K. Paliwal*[1]

[1]School of Engineering, Griffith University, Brisbane, QLD 4111, Australia

aadel.alatwi@griffithuni.edu.au, s.so@griffith.edu.au, k.paliwal@griffith.edu.au

Abstract

In this paper, we propose a new linear prediction (LP) analysis method for estimating LP coefficients that are used in current speech coders. This method improves the robustness of LP coefficients computed from noise-corrupted speech as well as enabling better quantisation efficiency. The study compares the quantisation performance, as well as the noise-robustness between the proposed LP coefficients and the conventional LP coefficients, in terms of spectral distortion. The results indicate that the proposed LP coefficients were more robust to noise and also enabled transparent coding at lower bitrates (savings of up to two bits/frame) than the conventional LP coefficients.

Index Terms: linear prediction coefficients, linear prediction analysis, split vector quantisation, spectral distortion

1. Introduction

The linear prediction (LP) analysis method is extensively used as a basic technique for low bitrate speech coding applications [1]. In these applications, LP coefficients, which represent the short-time power spectrum of a speech signal using a low-order all-pole filter [2], are typically obtained using the autocorrelation method [3], converted into LP parameters (e.g. line spectral frequency (LSF)) and then quantised using as few bits as possible for transmission [4]. In noise-free environments, the performance of LP parameter-based speech coders is often satisfactory. However, in the presence of background noise, the LP analysis method yields a poor estimate of the LP spectrum of the input speech signal; hence, the variance of the estimated LP coefficients is increased [5], which results in an overall deterioration in the reconstructed speech quality [6][7].

A pth all-pole filter $H(z)$ that is driven by white Gaussian noise on the input is used to represent the speech production model:

$$H(z) = \frac{G}{1 + \sum_{k=1}^{p} a_k z^{-k}} \tag{1}$$

The filter coefficients (also known as LP coefficients) a_k and filter gain G are estimated by solving the Yule-Walker equations:[1]

$$\sum_{k=1}^{p} R(j-k)a_k = -R(j), \quad \text{for } k = 1, 2, \ldots, p \tag{2}$$

$$G = R(0) + \sum_{k=1}^{p} a_k R(k) \tag{3}$$

where $R(k)$ are the autocorrelation coefficients estimated from a frame of N samples of the speech signal $x(n)$:

$$R(k) = \frac{1}{N} \sum_{k=0}^{N-1-k} x(n)x(n+k) \tag{4}$$

Another method of estimating autocorrelation coefficients utilises the Einstein-Wiener-Khintchine theorem, by taking the inverse discrete-time Fourier transform of the periodogram estimate of the power spectrum $P(\omega)$:

$$R(k) = \frac{1}{2\pi} \int_{-\pi}^{\pi} P(\omega) \mathrm{e}^{j\omega k} d\omega \tag{5}$$

$$\text{where} \quad P(\omega) = \frac{1}{N} \left| \sum_{n=0}^{N-1} x(n) \mathrm{e}^{-j\omega n} \right|^2 \tag{6}$$

This relationship between the periodogram and autocorrelation coefficients motivates our method for reducing the variance of the LP coefficient estimates, by smoothing the power spectrum before the LP analysis.

In this paper, we propose a modified procedure for performing LP analysis that reduces the variance of LP coefficient estimates in order to provide some noise robustness and also exploits the non-linear frequency selectivity of the human auditory system. The LP coefficients are fully compatible with current speech coding standards and the algorithm proposed can be easily incorporated into existing code implementations. We evaluate the quantisation performance of the proposed LP parameters using different quantisation schemes, as well as their robustness to noise in terms of the spectral distortion (SD). The results indicate that the proposed method provides a more accurate and noise-robust estimation of the LP parameters.

This paper is organized as follows: Section 2 describes the proposed LP analysis method. Section 3 outlines the experimental setup that we used to measure the quantization performance and the robustness of the LP parameters. Section 4 clarifies the details of each experiment and provides the results. Section 5 presents the conclusion of this study.

2. Proposed LP analysis method

The proposed LP analysis method uses two steps to compute the LP coefficients. The first step involves the manipulation of the periodogram of the input speech in order to reduce the variance of the spectral estimates and effects of noise. The second step applies the conventional autocorrelation method using the modified autocorrelation coefficients, which are computed by taking the inverse FFT of the processed power spectrum. The processed power spectrum is obtained through a smoothing operation using triangular filters that are spaced linearly on the Bark frequency scale [8]. These triangular filters simulate the human

[1] It can be readily shown that all-pole modelling is equivalent to linear prediction analysis. Therefore, we will refer to this estimation process as *LP analysis* from now on.

auditory system [9]. Also, there is a general downward spectral tilt in the speech power spectrum because there are more high energy formants located in the low frequencies and less energy peaks in the high frequencies. Therefore, the non-linear smoothing operation is performed less at low frequency regions and is performed more at high frequency regions [9], as shown in Figure 1.

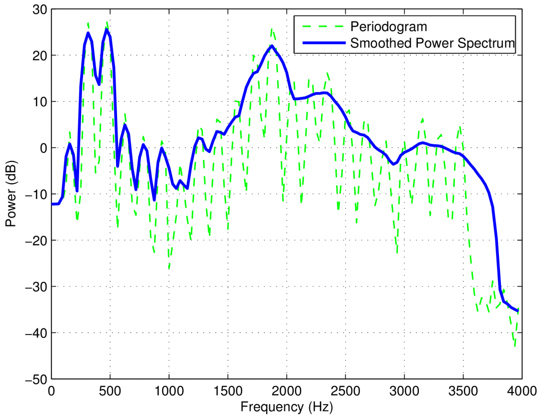


Figure 1: *Periodogram $P(k)$ and the smoothed power spectrum $\bar{P}(k)$ of speech sound (vowel / e / produced by male speaker).*

The proposed algorithm is described in the following steps:

Step 1: Compute the power spectrum $P(k)$ of a given frame $\{x(n), \quad n = 0,1,2,...,N-1\}$ of N samples from a speech signal using the FFT [10]:

$$P(k) = \frac{1}{N}\left|\sum_{n=0}^{M-1} x(n)w(n)\mathrm{e}^{-j2\pi kn/M}\right|^2, \quad 0 \leqslant k \leqslant M-1 \tag{7}$$

where $P(k)$ is the estimated power spectrum at the k^{th} normalized frequency bin, M is the FFT length where $M > N$, and $w(n)$ is a Hamming window.

Step 2: Determine the estimated smoothed power spectrum $\bar{P}(k)$ using a triangular filter at every frequency bin:

$$\bar{P}(k) = \sum_{i=-C(k)}^{C(k)} B(i)P(i-k) \tag{8}$$

where $B(i)$ is the triangular filter, which is spaced using the Bark frequency scale [8], and $C(k)$ is half the critical bandwidth of the triangular filter at frequency bin k.

Step 3: Take the inverse FFT of $\bar{P}(k)$ to compute the autocorrelation coefficients [10]:

$$\hat{R}(q) = \frac{1}{M}\sum_{k=0}^{M-1} \bar{P}(k)\mathrm{e}^{j2\pi kq/M} \quad , \quad 0 \leqslant q \leqslant M-1 \tag{9}$$

These autocorrelation coefficients $\hat{R}(q)$, $0 \leqslant q \leqslant p$, where p is the LP analysis order, are used in the Levinson-Durbin recursion algorithm [10] to compute the LP coefficients, which we called the Bark frequency Smoothed Linear Prediction (BS-LP) coefficients.

The behavior of the BS-LP analysis method in spectral modelling of speech signal is demonstrated in the example shown in Figure 2. In this figure, the periodogram $P(k)$ of a frame of the speech sound (vowel /e/) is shown together with the all-pole spectral models of order $p = 10$ that were computed with two techniques: the proposed BS-LP and the conventional autocorrelation LP analysis method. As illustrated in Figure 2, at high frequencies, the formants appear to have wider bandwidths due to the large critical bandwidths at these frequencies, where more smoothing is performed. Therefore, this added smoothing reduces the influence of noise components at higher frequencies.

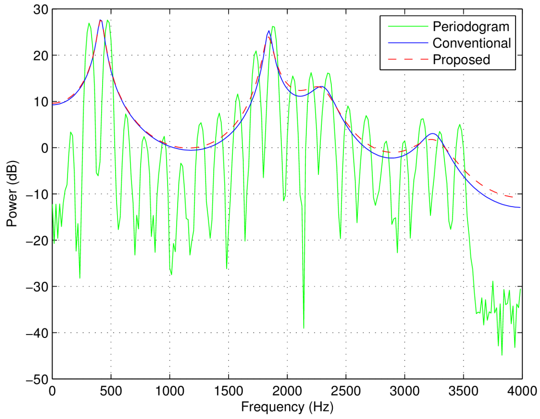


Figure 2: *Periodogram $P(k)$ with corresponding all-pole spectra of order $p = 10$ computed by the conventional LP and the proposed BS-LP analysis method of clean speech (vowel /e/ produced by male speaker).*

3. Experimental setup

3.1. Database

The TIMIT database was utilized for all of the simulations performed for this paper. It contains of 462 train speakers and 168 test speakers. The database was downsampled to 8 kHz. The estimation of the spectral envelope was carried out using the FFT length of 256 frequency samples. A 10^{th} order LP analysis was performed with high frequency compensation on a 20 ms analysis framework. A 10 Hz bandwidth expansion was applied because of sharp spectral peaks of LP spectrum that are caused by the underestimation of the formant bandwidths [4].

3.2. Performance evaluation criterion

In order to determine the quality of the power spectrum, the SD (spectral distortion) of the estimated spectral envelope was computed over the power spectrum of a frequency plane as an objective measure. It is defined as [4]:

$$SD = \sqrt{\frac{1}{F_s}\int_0^{F_s}[10\log_{10}P(f) - 10\log_{10}\hat{P}(f)]^2 df} \tag{10}$$

where F_s is the sampling frequency, and $P(f)$ and $\hat{P}(f)$ denote the true and estimated power spectra, respectively. As can be observed in Equation (10), a low SD indicates the reconstructed speech spectral envelope to be closer to that of the original speech, and therefore is of better quality. This distortion measure is carried out on the power spectrum produced from 20 – 30 ms frames of speech. The measure will be utilized to measure the accuracy and robustness of the proposed BS-LP analysis method in Section 4.1.

The number of bits allocated for quantisation influences the performance of the quantisation of the LP parameters. In many cases, the number of bits allocated for quantisation is found when the preferred rate of spectral accuracy has been achieved. This provides an equivalent basis of comparison between the proposed BS-LP analysis method and the conventional LP analysis method. To evaluate the quantisation process, the SD will be observed in two different classifications: the average SD for the whole data and the percentage of outlier frames. An outlier frame has an SD $\geq$ 2 dB. The outlier frames are divided into the following types: outlier frames with an SD in the range of 2 – 4 dB; and outlier frames with an SD greater than 4 dB. The preferred performance for the quantisation of the LP parameters is when transparent coding is achieved [4], which is defined by the following conditions:

1. The average SD is about 1 dB.
2. The number of outlier frames with an SD between 2 – 4 dB is less than 2%.
3. No outlier frames are greater than 4 dB.

4. Results and discussion

4.1. Noise robustness analysis

The robustness of the spectral envelope estimated using the BS-LP analysis method was compared with robustness found using the conventional LP analysis method. Simulations to find the robustness of the proposed method were carried out by measuring the SD between the power spectrum of the clean and noisy signal, respectively. The results that were obtained from this experiment are shown in Figure 3. The data shows that the BS-LP parameters appeared more robust to noise than the conventional LP parameters; the SDs of the BS-LP analysis method were consistently lower than the conventional LP analysis method for all Signal to Noise Ratios (SNRs). By referring to the example shown in Figure 2, this behaviour can be explained by the effectiveness of the smoothing operations in the computation of the BS-LP parameters.

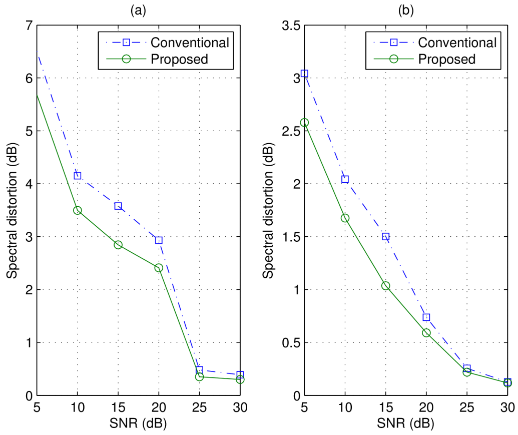


Figure 3: *Spectral distortion values (SD) between the conventional and proposed LP spectral envelopes of order $p = 10$ computed from clean and noisy speech (vowel / e / produced by male speaker). Speech was corrupted by two types of noise: (a) additive zero-mean Gaussian white noise, and (b) street noise, in six SNR categories.*

4.2. Quantisation performance of LP parameters

The full-search vector quantisation (VQ) has a high computational complexity that requires excessive memory space in order to perform the quantisation of the codebook. Though the split VQ mechanism is suboptimal, it lowers both the computational complexity and the required memory space to a manageable level without significantly affecting the performance of the VQ [11]. As a result, we used the split VQ to study the quantisation performance of the LP parameters.

In this split VQ, the LSF vector, which is a popular representation of the LP parameters, is separated into different parts of the lower order. The codebooks of the VQ were designed using the Linde-Buzo-Gray (LBG) algorithm [12] with the weighted LSF distance measure for every part, which is given by [4]:

$$d(f,\hat{f}) = \sum_{\mathbf{i=1}}^{\mathbf{10}} [c_i w_i (f_i - \hat{f}_i)]^{\mathbf{2}} \qquad (11)$$

where f_i and $\hat{f}_i$ are the i^{th} LSF representation in the approximated and original vector, respectively, and the weights w_i and c_i are assigned to the i^{th} LSF. The variable weight w_i is given by [4]:

$$w_i = [P(f_i)]^r \qquad (12)$$

where $P(f)$ is the LP power spectrum and r is equal to 0.15. The fixed weight c_i is given by [4]:

$$c_i = \begin{cases} 1.0 & \text{for } 1 \leq i \leq 8 \\ 0.8 & \text{for } i = 9 \\ 0.4 & \text{for } i = 10 \end{cases} \qquad (13)$$

We used two-split VQ (the first part had 4 LSFs and the second part had 6 LSFs) and three-split VQ (the first part had 3 LSFs, the second part had 3 LSFs, and the third part had 4 LSFs).

In all experiments, the quantisation performance for each method was evaluated using the SD measure, as given in Equation (10), where $P(f)$ and $\hat{P}(f)$ are the power spectrum of the original and reconstructed speech, respectively, and F_s covers the partial-band SD (i.e. 0 – 3kHz), which is used to evaluate the quantisation schemes that use a weighted distance measure [13].

The quantisation performances obtained from the two-split VQ experiments are listed in Tables 1 and 2 for the conventional LP and BS-LP analysis methods, respectively. The results indicate that the two-split VQ that uses the conventional LP analysis method required 24 bits/frame to achieve the transparent quantisation. However, the BS-LP analysis method required only 22 bits/frame, which saved 2 bits/frame from the conventional LP analysis method. The outlier frames percentage between 2 – 4 dB was much lower in favour of the proposed BS-LP analysis method.

Tables 3 and 4 illustrate the results obtained from the three-split VQ for the conventional LP and BS-LP analysis methods, respectively. It can be seen that the proposed BS-LP analysis method offers an advantage of 1 bit/frame over the conventional LP analysis method, which required 25 bits/frame to achieve the quantisation transparency.

5. Conclusion

This paper presented a modified method for estimating LP coefficients for current speech coders, by applying non-linear smoothing to the power spectrum. These BS-LP coefficients

Table 1: *Average SD of the two-split vector quantizer as a function of bitrate (using the conventional LP analysis method with the weighted LSF distance measure).*

Bits	Av. SD	Outliers (%)	
used	(in dB)	2 – 4 dB	>4 dB
26	0.90	0.39	0.00
25	0.95	0.58	0.00
24	1.04	1.07	0.00
23	1.09	1.68	0.00
22	1.19	3.16	0.00
21	1.26	4.23	0.00
20	1.31	6.10	0.00

Table 2: *Average SD of the two-split vector quantizer as a function of bitrate (using the proposed BS-LP analysis method with the weighted LSF distance measure).*

Bits	Av. SD	Outliers (%)	
used	(in dB)	2 – 4 dB	>4 dB
26	0.77	0.12	0.00
25	0.85	0.19	0.00
24	0.89	0.26	0.00
23	0.97	0.53	0.00
22	1.02	0.79	0.00
21	1.11	1.57	0.00
20	1.18	2.44	0.00

Table 3: *Average SD of the three-split vector quantizer as a function of bitrate (using the conventional LP analysis method with the weighted LSF distance measure).*

Bits	Av. SD	Outliers (%)	
used	(in dB)	2 – 4 dB	>4 dB
30	0.78	0.20	0.00
29	0.80	0.27	0.00
28	0.85	0.51	0.00
27	0.88	0.56	0.00
26	0.97	0.91	0.00
25	1.05	1.73	0.00
24	1.18	3.10	0.01
23	1.21	3.83	0.01
22	1.30	5.91	0.03

Table 4: *Average SD of the three-split vector quantizer as a function of bitrate (using the proposed BS-LP analysis method with the weighted LSF distance measure).*

Bits	Av. SD	Outliers (%)	
used	(in dB)	2 – 4 dB	>4 dB
30	0.69	0.09	0.00
29	0.74	0.14	0.00
28	0.81	0.35	0.00
27	0.83	0.38	0.00
26	0.90	0.54	0.00
25	0.97	1.28	0.00
24	1.03	1.38	0.00
23	1.12	2.04	0.00
22	1.25	4.53	0.02

possess improved robustness to noise. The advantage of this method is that it is fully compatible with current speech coder implementations. The LP quantisation performance of the BS-LP coefficients, in comparison with the conventional LP coefficients, was investigated for various split vector quantisation schemes. For both two- and three-split vector quantisation, the proposed BS-LP analysis method offered a saving of about 2 bits/frame and 1 bit/frame over the conventional LP analysis method, respectively, in terms of the spectral distortion. In addition, the results demonstrated the improved noise-robustness of the BS-LP coefficients for low to medium SNR levels and for both white and street noise. Since the Bark frequency triangular filters used for the power spectrum smoothing are similar to those used in the computation of Mel frequency cepstral coefficients (or MFCCs), these BS-LP coefficients may exhibit better recognition performance and noise robustness in automatic speech recognition tasks. We will investigate the use of BS-LP coefficients in the network speech recognition context in a future paper.

6. References

[1] J. Makhoul, "Linear prediction: A tutorial review," *Proceedings of the IEEE*, vol. 63, no. 4, pp. 561–580, 1975.

[2] B. S. Atal, V. Cuperman, and A. Gersho, *Speech and Audio Coding for Wireless and Network Applications*. Springer Science & Business Media, 2012, vol. 224.

[3] A. El-Jaroudi, "Linear predictive coding," *Encyclopedia of Telecommunications*, 2003.

[4] W. B. Kleijn and K. K. Paliwal, *Speech coding and synthesis*. New York, NY, USA: Elsevier Science Inc., 1995.

[5] S. M. Kay, "The effects of noise on the autoregressive spectral estimator," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 27, no. 5, pp. 478–485, 1979.

[6] M. R. Sambur and N. S. Jayant, "Lpc analysis/synthesis from speech inputs containing quantizing noise or additive white noise," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 24, no. 6, pp. 488–494, 1976.

[7] A. Trabelsi, F. Boyer, Y. Savaria, and M. Boukadoum, "Improving lpc analysis of speech in additive noise," in *Circuits and Systems, 2007. NEWCAS 2007. IEEE Northeast Workshop on*. IEEE, 2007, pp. 93–96.

[8] H. Fletcher, "Auditory patterns," *Reviews of modern physics*, vol. 12, no. 1, p. 47, 1940.

[9] B. Moore, *An Introduction to the Psychology of Hearing*. Academic Press, 1997.

[10] M. H. Hayes, *Statistical digital signal processing and modeling*. John Wiley & Sons, 2009.

[11] K. K. Paliwal and B. S. Atal, "Efficient vector quantization of lpc parameters at 24 bits/frame," *Speech and Audio Processing, IEEE Transactions on*, vol. 1, no. 1, pp. 3–14, 1993.

[12] Y. Linde, A. Buzo, and R. M. Gray, "An algorithm for vector quantizer design," *Communications, IEEE Transactions on*, vol. 28, no. 1, pp. 84–95, 1980.

[13] S. So and K. K. Paliwal, "Efficient product code vector quantisation using the switched split vector quantiser," *Digital Signal Processing*, vol. 17, no. 1, pp. 138–171, 2007.

Synthesizing Attitudes in German

Angelika Hönemann[1], *Petra Wagner*[1]

[1] Bielefeld University, Faculty of Linguistics and Literary Studies, Germany

ahoenemann@techfak.uni-bielefeld.de, petra.wagner@uni-bielefeld.de

Abstract

Our study investigates the potential of modeling the synthetic realization of four human attitudes (uncertainty, sincerity, surprise and doubt) based on a set of prosodic and voice quality parameters. A set of acoustic parameters were extracted from a corpus of German expressive speech. A comparison of lexically identical human and synthesized expressive utterances yields mostly positive correlations between the acoustic parameters used for analysis and modeling. In a subjective evaluation, listeners were asked to identify a target attitude in pairs of synthesized utterances. That way, uncertainty was identified in 90%, followed by sincerity (80%), surprise (72%) and doubt (64%).

Index Terms: speech synthesis, expressive speech, computational paralinguistics, prosody, voice quality

1. Introduction

High quality synthetic speech output is an indispensable attribute of any intelligent system such as a virtual agent or robot that uses speech-based communication when interacting with humans. Thus, speech synthesis research has begun to focus on the optimization of speech synthesis to fit the needs of speech-based Human-Machine Interaction (HMI) or dialogue systems [16]. Such interactions go beyond the mere exchange of words and/or factual information. In order to aid interactive grounding, comprehension and floor management, speakers use prosodic means to convey meta-information about the relevance, novelty and importance of what has been said, express information on their cognitive status (e.g. attention by feedback behavior), the ongoing speech planning process (e.g. by hesitations), floor management (e.g. by providing prosodic turn yielding cues) and emotions or attitudes related to the ongoing dialogue situation [15].

The expression of attitudes is a highly relevant factor in social interaction. Unlike emotions, they express the *affectively loaded cognitive appraisal* of a situation (or an object) [4]. We assume that the expression of attitudes can be of a short-timed, transitional nature and is likely to be ubiquitous in everyday communication. Hence, the expression of attitudes may be a crucial factor in HMI, making it more robust, as additional, nonverbal information is transported through the speech channel. E.g., a dialogue system could react to a low reliability of the speech recognition by expressing its subsequent reaction with the attitude of *uncertainty*, thereby implicitly making a confirmation request and critically reducing the number of necessary dialogue turns.

Previous studies have shown that attitudes are expressed through fine-grained adaptations of multiple acoustic prosodic and voice quality related parameters [5, 10]. Therefore, parametric rather than concatenative approaches to speech synthesis are probably suited best for its realization. However, to this day, dialogue systems tend to rely on concatenative approaches to synthesis such as unit selection or slot-and-filler systems, probably due to their high quality and because dialogue systems tend to operate in limited domains. It remains to be shown whether the potential benefits of attitudinal synthesis are strong enough to surpass the quality limitations introduced by parametric synthesis [8].

This paper presents a first feasibility study to explore the possibility of modeling and perceiving –often subtly expressed– attitudes with the help of adaptable parametric synthesis. In the remainder of this paper, we discuss results of an acoustic analysis of attitudinal expression based on German corpus data taken from previous work (section 2, [6]). In section 3, we describe the development of a set of rules for parameter adaptation in synthetic speech for four attitudes (sincerity, uncertainty, doubt, surprise). Section 4 describes the objective evaluation of the resulting attitudinal speech synthesis, section 5 presents the subjective evaluation based on a simple discrimination task. The paper closes with a discussion and a conclusion (section 6).

2. Data Analysis

The present analyses are based on a previously collected corpus of paralinguistic German speech [6]. The whole corpus consists of productions of two short utterances (*Marie tanzte*, *Eine Banane)* produced in 16 different attitudes by 20 native German speakers (11f., 9m). The full corpus contains a total of 640 utterance recordings. All utterances were force-aligned on phone level and SAMPA-transcribed using the Munich AUtomatic Segmentation system MAUS [12]. For each utterance, a set of acoustic parameters related to paralinguistic expression (F0, intensity, duration, jitter, shimmer) was extracted with [3]. The analyses showed that two of the selected attitudes are prototypically realized with a rising, "interrogative" contour (doubt, surprise), while two others tend to follow a falling, "declarative" contour (uncertainty, sincerity). These four attitudes were selected for further analysis and synthesis modeling. In total 80 stimuli (10 speakers * 4 attitudes * 2 utterances) were analyzed. Cross-speaker averages of these analyses are presented in Table 1 (human speakers).

3. Adaptable Synthesis

In order to realize the adaptation of synthetic speech according to the results of the acoustic analysis, a version of the MaryTTS system [14] embedded into the incremental speech processing system InproTK was used [1]. InproTK offers a 'just in time' modification of the speech parameters during the synthesis, thus, it can react immediately to dynamically changing situations during an ongoing discourse, e.g. those that require an attitudinal reaction. InproTK provides modules to realize modifications of the synthesis output [2], but these are limited to the HMM synthesis offered by MaryTTS.

3.1. Acoustic parameter matching

We used the attitude specific mean values across human productions as input for calculating phone durations, fundamental frequency (F0), intensity as well as voice quality (VQ) parameters such as jitter and shimmer, since these acoustic parameters have been shown to be crucial for the perception of different attitudes [5, 9, 10].

This initialization is performed on phone level, while distinguishing between the phone classes of *vowels (V). consonants (C)* and *long vowels (LV)*. Additionally, a random factor ranging from zero to the standard deviation of each feature (RF) was added to the mean of each respective feature. This random factor simulates the measured speaker-specific variations in the resulting synthesized productions across the various attitudinal states. For the initialization, we defined the position of each phone and computed the percentage (PF) of the overall mean of an acoustic feature either for a phone at the first, middle and last position in a word or utterance or of a stressed phone. This allows for marking of stressed positions and stress related lengthening.

MaryTTS was used to generate the relevant acoustic parameters (F0, intensity, duration, phone duration) used for common synthesis. The parameters were then adapted by equations 1-5, based on the attitude-specific means of the various speech parameters for each phone (i), based on human productions [6]. Furthermore, interdependencies between the various acoustic parameters – especially between F0 and intensity – were derived from our empirical analyses. From these, we derived a set of heuristic rules used in the synthesis modeling: The exact rules are described in the equations below. Additionally, correspondences between F0 and intensity were modeled by adding the attitude-specific variability of intensity on F0 and vice versa. This leads to an increase of F0 or intensity based on attitude-specific variability in the corresponding acoustic domain.

Phoneme duration (Dur) The duration generated by MaryTTS (gDur) is shifted by a factor based on the sum of the duration derived from the human analysis (setDur) and the random factor of the duration (RF) multiplied by the position of the phone (PF(i)) divided by 100.

$$Dur_{i=1}^{n} += \frac{(setDur + RF_{dur})\; PF_i}{100}\; gDur_i \qquad (1)$$

Intensity (Int) The intensity is based on the sum of the intensity derived from the human analysis (setInt) and the RF(s) of the intensity. The sum is multiplied by the phone's PF and added to the intensity.

$$Int_{i=1}^{n} += \; (setInt + RF_{int})\; PF_i \qquad (2)$$

Fundamental Frequency (F0) The F0 is the sum of the F0 derived from the human analysis (setF0) and the RF of F0. The sum is multiplied by the phone's PF and added to the F0.

$$F0_{i=1}^{n} += (setF0 + RF_{F0})\; PF_i \qquad (3)$$

Jitter (Jit) The jitter is the sum of the jitter derived from the human analysis (setJit) and the RF of the jitter. The sum is multiplied by the phone's PF and added to the jitter.

$$Jit_{i=1}^{n} += \left(setJit + RF_{Jit}\right) PF_i \qquad (4)$$

Shimmer (Shim) The shimmer is the sum of the shimmer derived from the human analysis (setShim) and the RF of the shimmer. The sum is multiplied by the PF of the phone and added to the shimmer.

$$Shim_{i=1}^{n} += (setShim + RF_{Shim})\; PF_i \qquad (5)$$

3.2. Adaptation process

The general adaptation process is shown in Figure 1. It starts with the initialization of the *AdaptableSynthesisModule.* This module implements each phone as the *SysSegmentIU* class of the utterance and determines its position in a word and utterance. Furthermore the utterance mode is assigned.

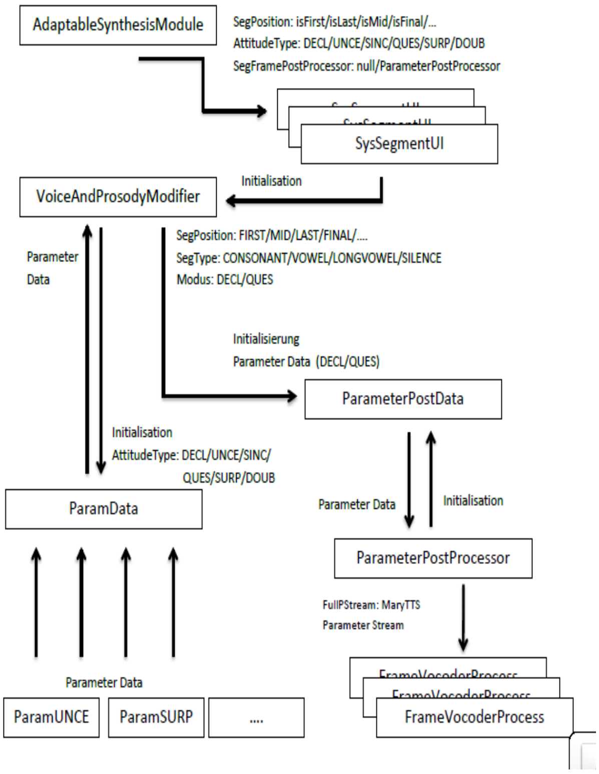


Figure 1: *Schematic diagram of synthesis process*

Each phone holds a class *VoiceAndProsodyModifier* including the equations for computing each feature value and sets the speech parameters for the current phone accordingly. This class receives the values for a specific attitude from the parameter class such as *ParamUNCE* for the uncertainty values or *ParamSURP* for the surprise values. All parameters are defined in these classes. Finally the *PostParameterData* container holds all relevant values for post processing. During post processing, the parameters of the MaryTTS HMM model for the common synthesis are adapted before the actual vocoding process starts. F0, intensity as well as the spectral information are computed for each frame. Each frame of a MaryTTS voice has a period of 5ms (200/sec).

Jitter and shimmer describe irregularities of the F0 (jitter) and the energy (shimmer) in the voice. The irregularity of F0 (cf. eq. 3) and intensity (cf eq. 2) are computed following the procedure in [9]: For each frame (i) we calculated a factor using the mean jitter derived from the human data (cf. eq. 4) as a multiplier to compute three sine waves, which are then added to each F0 value (cf. eq. 6, 7).

Table 1: *Means and standard deviation of the speech parameter for synthesized and spoken attitudes (across speaker and utterance) for five males (M) and five females (F)*

			Duration (ms)		F0 (Hz)		Intensity (dB)		Jitter (%)		Shimmer (%)	
		attitudes	mean	sd	mean	Sd	mean	sd	mean	sd	mean	sd
Human Speaker	M	sincerity	84	41	103.49	23.74	73.12	6.11	2.75	1.83	14.79	7.70
		uncertainty	102	58	96.65	20.94	70.32	6.53	3.00	2.43	12.99	8.89
		doubt	112	77	111.98	50.11	71.23	6.15	3.19	2.13	16.35	9.64
		surprise	115	75	127.10	58.53	72.84	6.43	3.59	2.86	13.14	6.32
	F	sincerity	94	61	188.75	38.11	72.60	8.10	2.91	2.35	13.88	7.59
		uncertainty	107	83	192.75	43.68	70.88	7.48	2.75	2.56	12.76	7.69
		doubt	112	71	196.61	65.75	70.39	6.51	3.53	3.08	15.40	10.20
		surprise	111	71	216.06	74.74	71.06	7.35	3.85	3.23	13.60	6.84
Synth. Adp. MaryTTS	M	sincerity	149	118	111.13	14.71	54.81	8.60	1.25	0.79	6.81	2.71
		uncertainty	160	160	110.21	14.97	54.75	9.29	1.47	2.15	7.24	5.47
		soubt	133	107	118.12	14.60	55.59	7.83	1.51	0.99	8.00	4.79
		surprise	121	85	117.84	14.48	55.10	8.58	1.43	0.88	7.42	3.78
	F	sincerity	149	120	150.72	21.95	63.67	8.05	1.27	1.15	8.20	5.32
		uncertainty	168	141	150.49	21.69	63.01	8.43	1.25	1.12	7.97	4.77
		doubt	189	163	172.78	15.63	64.97	8.57	1.42	1.85	7.52	5.73
		surprise	148	114	173.43	15.80	65.32	8.53	1.42	2.35	7.26	6.72

We used the same process for the intensity adaptation. A multiplier is computed using the shimmer yielded from the human data (cf. eq. 5, 9) to calculate the sine waves added to each energy value (cf. eq. 10). Finally each F0 and energy value within a frame is subtracted from the current mean of the phone to ensure smooth transitions (cf. eq. 8, 11).

$$FJ_{i=1}^{n} = setJit\ 100\ \pi\left(\frac{i}{200}\right) \qquad (6)$$

$$F0_{i=1}^{n} \mathrel{+}= \sin(12.7FJ) + \sin(7.1FJ) + \sin(4.7FJ) \qquad (7)$$

$$F0_{i=1}^{n} = \emptyset F0 - F0_i \qquad (8)$$

$$FS_{i=1}^{n} = setShim\ 100\ \pi\left(\frac{i}{200}\right) \qquad (9)$$

$$Eng_{i=1}^{n} \mathrel{+}= \sin(12.7FS) + \sin(7.1FS) + \sin(4.7FS) \qquad (10)$$

$$Eng_{i=1}^{n} = \emptyset Eng - Eng_i \qquad (11)$$

The vocoding process produces an audio stream on a frame-by frame-basis until the whole utterance is finished or the vocoding process is interrupted. The audio stream can be heard immediately, i.e. adaptation is simultaneous to the voice output.

4. Objective Evaluation

To compare the result of the attitudinal synthesis adaptation with the human productions, we synthesized a set of utterances directly comparable with the human data used in the analysis (5f, 5m, simulated by the random factor). We then extracted the identical speech parameters from the synthetic productions as in the analysis of the human productions. Table 1 shows the means and standard deviations of the two utterances produced by 10 human speakers and their synthetic counterparts. As there was no significant difference between individual utterances, means were calculated across utterances.

In most cases, the synthesized acoustic parameters for males and females are smaller than their corresponding human parameters. An exception to this is duration, i.e. synthetic speech tends to be slower. The following differences between the analyzed acoustic parameters for males (M) and females (F) can be observed: Δ Dur_M=37.5, Δ Dur_F=57.5, Δ F0_M=9.15, Δ F0_F=36.7, Δ Int_M =16.8, Δ Int_F=6.9, Δ Jitter_M=1.7, Δ Jitter_F=1.9, Δ Shimmer_M=6.9, Δ Shimmer_F=6.2.

In order to get an estimate of the similarity between human and synthesized attitudes, we calculated correlations between two versions. For each acoustic parameter correlations are based on the mean values of each phone for both utterances (cf. Table 2). The tests yield high positive correlations for the majority of parameters, but a few marginal or even negative correlations in a few cases (displayed in red) indicate less fitting synthetic realizations.

Table 2: *Correlation coefficient for females (F) and males (M) for duration. F0, intensity, jitter and shimmer*

		Dur	F0	int	jitter	shim
M	sincerity	.74	.34	.78	-	.30
	uncertainty	.60	-	.71	.68	-.41
	doubt	.55	.75	.76	.37	-.09
	surprise	.68	.74	.79	.72	.41
F	sincerity	.47	.59	.74	.21	.19
	uncertainty	.32	-.55	.82	.44	-.07
	doubt	.86	.51	.79	.33	.27
	surprise	.83	.54	.75	-.12	.63

5. Subjective Evaluation

As the acoustic identification of attitudes is a difficult task even in human speech [6, 7, 11], a simple identification task was set up to assess the potential suitability of our approach. The evaluation was carried out with ten native German participants (5m, 5f). Each participant was asked to identify a (textually represented) target attitude out of a pair of two synthetic utterances representing different attitudes. A major discriminating feature of attitudes appears to be the global F0 contour (rising “interrogative”: *doubt/surprise*; falling “declarative”: *uncertainty, insecurity*). To exclude this all too obvious feature and to ensure that listeners need to take into account more subtle cues, only “interrogative” or “declarative” attitudes were compared with each other, i.e. “doubt vs. surprise” and “uncertainty vs. sincerity”. Participants were allowed to listen each stimulus repeatedly. The utterance *Diese Banane ist gebogen* (engl. *This banana is* *bent*) was synthesized with a male and a female voice for each of the four target attitudes and in five variations, using the random factor implemented in the synthesis strategy (see section 3). The variations simulate individual speaking styles. In total, our evaluation set contained 40 stimuli for identification, which were presented to listeners in 20 pairs. The stimulus pairs were assigned randomly within "interrogative" and "declarative" attitudes. The target attitude to be identified for each pair was selected randomly as well.

5.1. Results

The test yielded 50 identifications for each target attitude. Figure 2 shows the identification score (%) for each target attitude across subjects. Each participant identified the target attitudes better than chance level of 50%. Declarative attitudes were more convincing than interrogative ones. The best identification was found for uncertainty (90%, 45 of 50), the worst for doubt 64% (32 of 50), sincerity is identified in 80% (40 of 50) and surprise in 72% (36 of 50) of the cases.

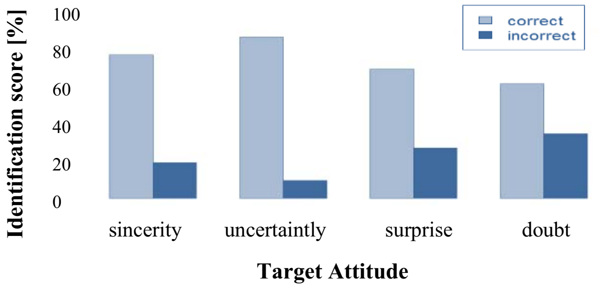


Figure 2: *Correct (lightblue) and incorrect (darkblue) identification score [%] of the synthesized attitudes across the subject*

6. Discussion and Conclusion

The current paper engaged in the parametric synthesis of four attitudinal states in German. Human recordings of attitudes have provided the empirical base for our synthesis strategy. We chose a rule-based approach because it offers a simple environment to identify and optimize the relevant parameters for an attitudinal synthesis and allows for a straightforward phonetic interpretation. The current work is a preliminary step for the later development of a model-based synthesis.

The usage of the unmodified results of the human analysis leads to a satisfactory simulation of the attitudinal states despite the obvious limitations of the HMM synthesis. The objective evaluation found that acoustic prosodic and voice quality parameters resemble those of the human originals. Simulating individual characteristics by introducing a random factor proved a successful approach.

The results of the subjective evaluation revealed that attitudes produced with a "declarative" contour were identified better than those with an "interrogative" contour. For now, we assume that the reason for this lies in the comparative proximity of *surprise* and *doubt* in function, form and their position in affective space [13]: *Surprise* and *doubt* share a rather high emotional activation, which has been shown to increase both F0 and intensity, while the declarative attitudes are more dissimilar: Uncertainty has a negative valence, while sincerity is considered as neutral. Furthermore, our results corroborate findings on the perception of attitudes expressed by human speakers, which have likewise shown that *doubt* and *surprise* can be less reliably identified in the absence of additional visual cues, i.e. facial expression [6, 7]. We therefore conclude that a less ambiguous synthesis of attitudinal behaviour needs to follow a multimodal approach.

7. Acknowledgements

This research was supported by the Cluster of Excellence Cognitive Interaction Technology 'CITEC' (EXC 277) at Bielefeld University, which is funded by the German Research Foundation (DFG).

8. References

[1] Baumann. T.. & Schlangen. D. The InproTK 2012 Release. In Proceedings of NAACL-HLT. 2012

[2] Baumann T. Schlangen D. INPRO iSS: A Component for Just-In-Time Incremental Speech Synthesis. In: Proceedings of the ACL 2012 System Demonstrations. ACL: 103–108.. 2012

[3] Boersma, P., Weenink, D:: Praat: doing phonetics by computer [Computer program]. Version 5.3.51, retrieved 2 June 2013 from http://www.praat.org/, 2013

[4] Fazio, R.H & M.A. Olson. Attitudes: Foundations, Functions, and Consequences. M.A. Hogg & J. Cooper. The SAGE Handbook of Social Psychology (pp. 139-160), London: Sage, 2003

[5] Gobl C., Chasaide A.N., The role of voice quality in communicating emotion, mood and attitude. Speech Communication 40, p. 189–212, 2003

[6] Hönemann. A.. Mixdorff. H.. Rilliard. A.. Social Attitudes - Recordings and Evaluation of an audio-visual Corpus in German. 7th Forum Acusticum. Krakow. Polen. 2014

[7] Hönemann, A., Rilliard A., Mixdorff, H., Classification of Auditory-Visual Attitudes in German, FAAVSP - The 1st Joint Conference on Facial Analysis, Animation and Auditory-Visual Speech Processing, Vienna, Austria 2015

[8] Huang, X., Acero, A. & H. Hsiao-Wuen. Spoken Language Processing. A guide to Theory, Algorithm, and System Development. Upper Saddle River, New Jersey: Prentice Hall.

[9] Klatt. D. & Klatt. L. Anaysis. synthesis. and perception of voice quality variations among female and male talkers. J. Acoust. Soc. America 87(2), 820-857, 1990

[10] Mixdorff. H.. Hönemann. A.. Rilliard. A. Acoustic-prosodic Analysis of Attitudinal Expressions in German. Proceeding of Interspeech 2015. Dresden. Germany. Page 1294, 2015

[11] Ricci B., Pio E., Luisa B., Paolo M., Roberto C., Pierluigi G., Expression and communication of doubt/uncertainty through facial expression. Journal of Theories and Research in Education, Ricerche di Pedagogia e Didattica, pp. 159-177, 2016.

[12] Schiel F, A statistical model for predicting pronunciation.. In: Proc. of the International Conference on Phonetic Sciences, Glasgow, United Kingdom, Paper 195, 2015

[13] Schauenburg, G., Ambrasat, J., Schröder, T., von Scheve, C., & Conrad, M. Emotional connotations of words related to authority and community. *Behavior Research Methods, 47*, 720-735, 2015

[14] Schröder. M. & Trouvain. J. The German Text-to-Speech Synthesis System MARY: A Tool for Research. Development and Teaching. International Journal of Speech Technology. 6. pp. 365-377. 2003

[15] Wagner, P. (What is) the contribution of phonetics to contemporary speech synthesis (?) D. Mehnert et al. (eds.). Systemtheorie, Signalverarbeitung, Sprachtechnologie. Studientexte zur Sprachkommunikation, Band 68 (pp.75–81.), TUD Press, Dresden, 2013.

[16] Ward, N. The challenge of modeling dialog dynamics. Workshop on Modeling Human Communication Dynamics at the 24th Annual Conference on Neural Information Processing Systems, Whistler, British Columbia, 2010.

Exploring Text To Speech Synthesis in Non-Standard Languages

Jesin James[1], *Catherine I. Watson*[1], *Deepa P. Gopinath*[2]

[1] Dept. of Electrical and Computer Engineering, University of Auckland, NZ
[2] College of Engineering, Trivandrum, India

Abstract

The development of Text to Speech synthesis systems as part of the Human Language Technology is a dominating research field in the present age. Languages are the backbone of every culture and developing technologies in every language is a basic requirement for users. Technology development focusing on the standard languages is a discouraging trend for many users who are not well-versed in these languages, but are in need of assistive technology. This paper focusses on the challenges in developing speech systems for non-standard languages, in light of a TTS development work in Malayalam language (non-standard language used in South India).

Index Terms: Under-resourced, Non-standard languages, Text To Speech synthesis, Indian languages, Malayalam

1. Introduction

Speech is the most basic form of human interaction and it has dominated all human ways of communication. This primary communication method was structured and categorized depending on the location where people lived, the differences in the words, tones, accents spoken etc. This gave birth to languages as we know them today. Language is the capability to convey ones emotions and ideas to a larger public. History has proved that the understanding of the scripts in ancient languages have paved way to envisage the development of the world as we see it today. In [1], S. Houston emphasizes this by stating Humankind is defined by Language; Civilization is defined by Writing. UNESCO has highlighted language diversity as a crucial element of the cultural diversity of the world [2][3].

Citing from the statistics depicted in Ethnologue website, presently there are a total of 7097 living languages in the world. Out of this, only 572 (8%) languages come under the category of "institutional", meaning they are used and sustained by institutions beyond the home country. Further analyzing the status of languages across the globe, a very discouraging fact emerges that only less than 100 (1.4%) languages have the required resources for high level language technology like sufficient speech corpus, parsers, POS taggers, morphological analyzers etc. [4] This is the era of advanced research studies in Human Language Technology including Text To Speech synthesis, Speech Recognition, Machine Translation, Natural Language Understanding etc. But it is a fact that a large majority of these works are happening only in a few privileged languages (1.4%) and the fruits of these research works do not reach people of all linguistic backgrounds. Such languages that are not having the necessary technical support to develop various speech processing technologies are termed as under-resourced languages or non-standard languages. There have been many distributed attempts[8] across the globe to analyse various non-standard languages and the predominant handicap reported by almost all researchers has been the non-availability of standard resources. The Basic Language Resource Kit (BLARK), a concept defined by Krauwer in [5] lists out the basic requirements for Text To Speech development in any language. BLARK comprises of written language corpora, spoken language corpora, mono and bilingual dictionaries, terminology collections, grammars, modules (e.g. taggers, morphological analysers, parsers, speech recognizers, text-to-speech), annotation standards and tools etc. Also there are many languages in the world where communication happens by speech only, as they do not have an acceptable written script. This makes speech processing complex [6].

The paper is organized as follows: Section 2 deals with the motivation in pursuing research in TTS systems in under-resourced languages. This is followed by Section 3 which explains various works that happened in other non-standard languages and the challenges faced by the researchers. Section 4 and 5 describes the background and methodology of the TTS development attempt in Malayalam language. Section 6 deals with the results that were obtained and Section 7 concludes the paper.

2. Motivation

Advancements in Human Language Technology is happening on a daily basis. It definitely makes our life easier and faster. But the real beneficiaries of these developments should be people who are vocally or visually challenged [8]. Statistics show that there are about 285 million visually impaired worldwide [9] out of which 90% are from developing countries. Also about 7% of the children in the age group of 13-17 face speech or language disorders [10]. These people are well distributed among all linguistic backgrounds, but the problems are more pronounced in developing and under-developing countries due to the lack of modern facilities. The TTS systems can be used by such people for book, newspaper reading and even for disabled children as an aid in studies. There is a growing awareness among disabled people on how these systems can make their life more simplified, but language becomes a barrier here as well.

Many studies show that cultural and linguistic backgrounds are dominating factors in acceptance of speech technology by people. [11] A sense of familiarity with a voice always encourages the people to use a particular technological system. A majority of the well-developed TTS technologies are in English, and there is a trend for non-English speakers to not accept them [12]. Especially for people with disabilities, learning a new language other than their mother tongue, to use a TTS system will be burdensome. Hence a necessity arises to develop high quality TTS systems in every non-standard language, for the ease of use of people and preventing the language from being extinct at a later stage. When such an attempt is made in minor languages, the researchers face the problem of inadequate resources as cited in the next section.

3. Related Work

One of the pioneering works in developing prosodic models in English is reported by Hirshberg et.al [32] in developing intonational phrasing using CART. In the experiment design for Australian aboriginal language Pitjantjatjara [13], the author has stated that developing a proper TTS for the language is not realizable presently, due to the non-availability of any resources for the same. So, the work reports using a major-language TTS to design a TTS for Pitjantjatjara. In [8] the author states that for a low-resourced language like Vietnamese, speech processing services or commercial products (ASR, TTS) do not exist yet. For developing such services, large amount of resources listed previously [16] are required. Due to this, cross lingual acoustic modelling has been used to develop a TTS for the language.

The authors in [14],[15] have reported TTS development and prosody incorporation research work in Yoruba and Ibibio which are languages spoken in Nigeria, Africa. They are both tone languages and hence prosody modelling is of utmost importance. Yoruba is spoken by about 28 million people, and Ibibio is being spoken by about 2 million people. But even then not much has been done to build computational resources for them. For the 171 living Philippine languages, there is no standard database available [17]. Due to this the researchers working in the language have to travel to the locality where the language is spoken, collect data from native speakers and then continue the research. There is an ongoing project to collect the corpora (Philippine Languages Online Corpora-PLOC), but it has been completed only for 8 languages. Even though the internet contains a plethora of text from many languages, a majority of world languages are not well represented in it. The Amharic language spoken in Ethiopia is one such language. An effort to build relevant corpus for Amharic has been carried out by Pellegrini et.al [18]. Celtic languages are Irish, Welsh, Breton, Scottish Gaelic, Cornish and Manx which are used in some parts of Europe. In [19], Delyth describes the difficulty in developing speech technology in these languages due to the non-availability of standard speech corpus. There are no newspapers and broadcast radio availability is also very limited. Only source of written speech data is from the internet, which are mainly translations from English. All this prevents any significant research for a TTS system in Celtic language

India has 22 official languages which are widely used in different parts of the country, but they are all low resource languages. An effort was taken by Hemant et.al. [21] to develop speech corpus for 13 of these languages. In the work reported in [22], a Kannada language TTS system is developed in Festival framework by first converting Kannada text to English and then using the TTS system. The work in Marathi language also reports the requirement of a larger and more standardized database to develop TTS system with good naturalness in the output speech [20]. There is a lack of standard databases in Indian languages. Efforts are being taken by various researchers, but they all have to be synchronized for working towards a common goal.

By investigating the status of Text To Synthesis in minor languages, it is very evident that there is a predominant “language divide” [7] in terms of the significant technological advancement and reachability to all the people in need. To bridge this divide, the only solution is to encourage research in these languages and attempt to develop standard resources for all languages. This has motivated the authors of this work to conduct prosody studies in Malayalam, a low-resourced Indian language.

4. TTS in Malayalam Language

Malayalam is one of the 22 official languages in India with 38 million native speakers particularly in the state of Kerala and the Laccadive Islands. There have been some attempts to develop TTS in Malayalam language. In [23] attempt is made to model the duration pattern of Malayalam speech in Festival speech engine. [24] reports using cluster analysis of duration patterns to improve the quality of Malayalam TTS. Arun et.al. explains an attempt to implement a concatenative synthesis method using ENSOLA technique [25]. But these developed systems and analysis have not produced TTS systems with the required speech quality and naturalness. This is because there is paucity in prosodic modeling, non-availability of standard databases, lack in enhancement of database to handle bilingual text for speech generation etc.

Prosody is defined as patterns of intonation and stress in a language and the various prosodic features are pauses, syllable prolongations, overall timing structure etc. The human brain introduces these features into natural speech to make it more understandable to the listeners; so, the inclusion of these prosodic features into synthesized speech will be similar to mimicking the human brain. The prosodic feature that is investigated and implemented in this work is “Pause”. Pause is the temporary stop introduced between words, phrases, sentences or paragraphs in speech. The inclusion of pauses in speech generally has two motives: one is to take sufficient breath to speak further and second is to provide the listeners’ sufficient time to decipher what was spoken. Pause is comprised of two parameters: pause position (where the pause has to be inserted in a sentence) and pause duration (for how long the pause has to be inserted). Modeling these two parameters is called Pause Modeling in speech and incorporating these pause patterns correctly will improve the overall duration model of synthesized speech.

Pause analysis has been conducted for other languages [26][27][28] like Mandarin Chinese, Japanese and Bangla, but pause patterns vary with language and speaking styles. Therefore to develop pause model for Malayalam language, speech corpus in Malayalam has to be analyzed in detail to derive the pause duration model. There has been no previous reported work in pause modeling for Malayalam. Since Malayalam is an under-resourced language, the availability of resources and standard speech corpus to suit the requirements of the study are limited. Hence such resources have to be collected and some have to be even recorded to complete the research work.

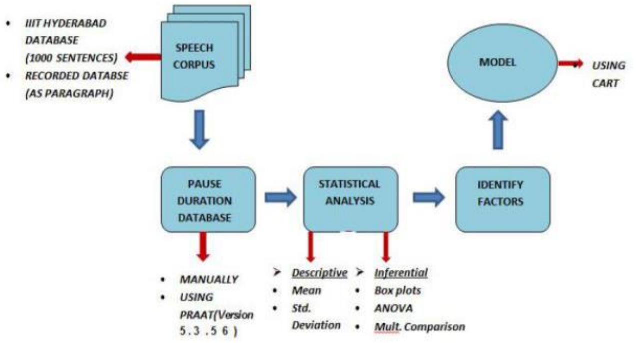


Figure 1: Methodology for pause analysis

5. Methodology

The basic methodology adopted for Pause Modelling of Malayalam TTS has been depicted in Figure 1. Getting an adequate speech corpus for analysis is the primary task. Since Malayalam is a non-standard language, the availability of a variety of databases that suit the needs of this particular study is limited. A database of 1000 Malayalam sentences compiled by IIIT-Hyderabad is used as Database I [29]. This database has 1000 .wav file recordings of individual sentences. Since the database has only individual sentences, it is inadequate for the analysis of pauses after sentence and after paragraph. So a new database (Database II) is recorded for the purpose which consists of 15 minutes (760 words, 7 paragraphs) of Malayalam speech read by a female Malayalam speaker. To analyse pauses, a pause duration database is required, and it is not existing for the language. So, a pause duration database is manually developed using the Praat speech analysis tool. Each of the sentences from the two databases are read in Praat and the pauses are manually marked and their durations measured to form the database. Along with each pause, its various parameters like location of the pause, the words/syllables before and after pauses and breath groups are also marked to form a consolidated database.

Once the pause duration database is developed, statistical analysis tools like mean, standard deviation (Descriptive analysis), box plots, ANOVA and Multiple comparison (inferential analysis) tools are employed. This analysis helps identify the various factors that affect pauses in Malayalam language. Once these factors are identified, a pause duration model is developed and an implementation of the model on Malayalam sentences is also conducted.

6. Results and Discussion

An attempt is made to identify the various factors that affect pauses in sentences. From the discussion with a linguist and reference to previous research work in other languages, the first factor studied is the Position of pause. The pause duration databases I and II are categorized in terms of the pause position and analyzed. The Database I has only individual sentences, so analysis of pause after word, phrase and comma are only possible. Database II consists of speech in the form of paragraphs, so along with the pause after word, phrase and comma; the analysis of pause after sentence and paragraph can also be conducted. The analysis tools used are box plots and ANOVA. In the box plots obtained, the medians of the boxes pertaining to each position did not fall on the same line and the overlap between the boxes is also minimal (Seen in Figure 2). The significance of Position of Pause as a factor for modelling can be further substantiated by the results from ANOVA analysis ($F_{3,280}$ = 283.2, $p < 0.05$). This clearly establishes that the pause at different positions have values in different ranges, and hence can be used as a factor for modelling pauses.

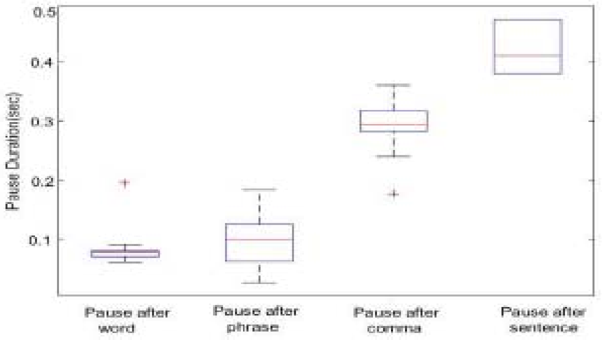


Figure 2: Box plot based on pause position (Database I)

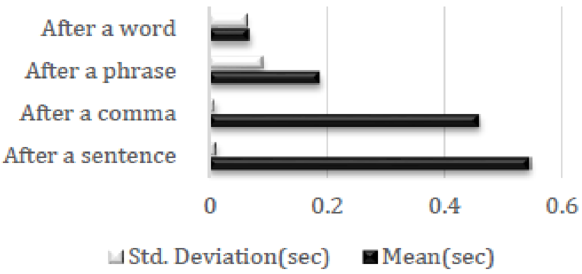


Figure 3: Descriptive analysis of pauses

Based on the position of pause as the primary factor, a general descriptive analysis of the pauses that occur at various positions in a sentence is conducted and depicted in Figure 3.The preliminary analysis suggests that the pause after word and phrase have high deviation values, as a result, further factors have to be identified to model them. But, pause after comma and sentence have low deviation values. So, they can be directly modelled by their mean values. [30][31]

The pause after phrase and word are investigated further by studying the databases. From inspection and discussions, more factors that affect these pauses are recognized (listed in Table I). To finalize these factors, statistical analysis tools like box plots, ANOVA were used.

Table I: Factors identifies for pause modelling

No.	Factor
1	Position of pause
2	Number of syllables before pause
3	Number of syllables in the word preceding the pause
4	Number of words before pause
5	Number of syllables after pause
6	Number of syllables in the word succeeding the pause
7	Number of words after pause

Pause Duration Model and Statistical Testing:
The identified factors are then used to model duration of pause after word and phrase using CART (Classification And Regression Tree) in MATLAB. Since the standard deviation values are low, the pause after comma and sentence are modelled directly by their mean values: 0.546s and 0.46s respectively.

Thus a complete model for pause is developed for Malayalam language.

In order to understand the effectiveness and accuracy of the model, RMSE (Root Mean Square Error) and Correlation tests are conducted. 60% of the database is used for training and 40% for testing. The developed model is implemented on the output of an existing Malayalam TTS system (in eSpeak), and the RMSE obtained is 0.025s with 90.85% Correlation in comparison with the testing corpus. This proves that the developed model predicts pause duration effectively.

7. Conclusion

An attempt is made to develop a Pause duration model for an under-resourced language. This is the first work reported in this regard for Malayalam language. Parts of the work have been reported in [30][31], but this paper focusses on the challenges in doing such a work for the first time in a non-standard language. Conducting research studies to develop TTS in under-resourced languages like Malayalam is a challenge, due to the many reasons emphasized in the Section 2 and 3 of the paper. If a user-friendly and high quality system is developed in a language, there will definitely be many users for it, especially physically challenged and aged people who prefer to communicate in their local language. A plethora of research work is happening in standard languages like English, but this is useful for only a minority of people who are well-versed in these languages. This leaves a very wide population still challenged because of their physical disabilities, even though there is such a huge advancement in technology. Also, researchers who attempt to work in the field of non-standard languages are often faced with the encumbrances of rarity in standard resources, difficulty in gathering the required resources, non-availability of written material in the language, inconvenience of traveling to the particular region to learn about the language etc. Researchers should be enthusiastic to work in their own languages and a Text To Speech technology accessible to all without any language divide should be the focus of future research.

8. References

[1] The First Writing, Script Invention as History and Process,S.Houston, ed. Cambridge Press, 2008.

[2] Jean-Marie Favre, Dragan Gasevic,Ralf La¨mmel,AndreasWinter, "Guest Editors' Introduction to the Special Section on Software Language Engineering",IEEE Transactions OnSoftware Engineering, December 2009

[3] UNESCO Ad Hoc Expert Group on EndangeredLanguages, "Language Vitality and Endangerment "

[4] Kevin P. Scannell "The Crúbadán Project: Corpus buildingfor under-resourced languages" in the Proceedings of the 3rdWeb as Corpus Workshop

[5] Steven Krauwer "The Basic Language Resource Kit(BLARK) as the First Milestone for the Language ResourcesRoadmap" in the Proceedings of SPECOM 2003

[6] L. Besacier, B. Zhou,Y. Gao, "Towards speech translationof non written languages," in Proc. SLT Workshop, Aruba,2006, pp. 222–225.

[7] Laurent Besacier , Etienne Barnard , Alexey Karpov , TanjaSchultz, "Automatic speech recognition for under-resourcedlanguages: A survey ", Speech communication, 2014

[8] T Dutoit, "An Introduction to Text-to-Speech Synthesis",Kluwer Academic Publishers.

[9] Statitics from IAPB(The International Agency forthe Prevention of Blindness)

[10] CDC/NCHS, National Health Interview Survey, 2012

[11] Rie Tamagawa ,Catherine I. Watson, I. Han Kuo ,Bruce MacDonald,Elizabeth Broadbent,"TheEffects of Synthesized Voice Accents on User Perceptions of Robots",International Journal of Social Robotics , 2011

[12] Jane M. Carey, Charles J. Kacmar "Cultural andLanguage Affects on Technology Acceptance and Attitude: Chinese Perspectives" International Journal of Information Technology,2010

[13] Harold Somers," Faking it: Synthetic text-to-speechsynthesis for under-resourced languages – Experimental design"ACL Anthology"

[14] Odéjobí, O.A., Wong, S.H.S., Beaumont, A.J.,"A modularholistic approach to prosody modelling for Standard Yorubaspeech synthesis" , Computer Speech and Language, 2008

[15] Ekpenyong, M., Udoh, E., Udosen, E., Urua, "Improvedsyllable-based text to speech synthesis for tone languagesystems", Lecture Notes in Computer Science, 2014

[16] Viet-Bac Le and Laurent Besacier, "Automatic SpeechRecognition for Under-Resourced Languages: Application toVietnamese Language" IEEE Transactions On Audio, Speech,And Language Processing, 2009

[17] Shirley Dita , Rachel Edita O. Roxas," PhilippineLanguages Online Corpora: Status, issues, and prospects", ACLAnthology

[18] T. Pellegrini and L. Lamel, "Investigating automaticdecomposition for ASR in less represented languages," in Proc.ICSLP'06, Pittsburgh, PA,2006

[19] D. Prys, "The BLARK Matrix and its relation to thelanguage resources situation for the Celtic languages," in Proc.LREC'06, Genova, Italy,2006.

[20] Mr.S.D.Shirbahadurkar,Dr.D.S.Bormane,"Marathi Langu- ge Speech Synthesizer Using Concatenative Synthesis Strategy(Spoken in Maharashtra, India)", Second InternationalConference on Machine Vision , 2009

[21] Hemant A Patil, "A Syllable-Based Framework for UnitSelection Synthesis in 13 Indian Languages", OrientalCOCOSDA ,2013

[22] Anusha Joshi, Deepa Chabbi , Suman M and SupritaKulkarni, "Text To Speech System For Kannada Language"IEEE ICCSP, 2015

[23] Bindhu;V. Rijoy;Deepa;Nimmy"Duration modeling for text to speech synthesis system using festival speech engine developed for Malayalam language", (ICCPCT), 2015

[24] K. S. Sreelekshmi ; Deepa P. Gopinath "Clustering ofduration patterns in speech for Text-to-Speech Synthesis" ,IEEE INDICON, 2012

[25] Arun Gopi,Shobana Devi Sajini,Bhadran, "Implementationof malayalam text to speech using concatenative based TTS forandroid platform" , IEEE ICCC, 2013

[26] Hiroya Fujisaki, Sumio Ohno, Seiji Yamada, "Analysis Of Occurrence Of Pauses And Their Durations In Japanese Text Reading", The 5th International Conference on Spoken Language Processing December, 1995

[27] Sudipta Acharya, Shyamal Kr.Das Mandal, "Occurrence And Duration Modeling Of Sentence Medial Pause For Bangla Text Reading At Different Speech Rate", in proc. of (Oriental COCOSDA), IEEE,2012

[28] Jian Yu , Jianhua Tao , "The Pause Duration Prediction for Mandarin Text to- Speech System", in proc. of IEEE NLPKE,2005

[29] Kishore Prahallad, E.Naresh Kumar, Venkatesh Keri,S.Rajendran, Alan W Black, "The IIIT-H Indic Speech Databases," INTERSPEECH, 2012

[30] James, J., Gopinath, D.P.,"Modeling pause duration for Malayalam language TTS",IEEE ICALIP, 2014

[31] Jesin James, Deepa P. Gopinath, "Pause Duration Model for Malayalam language TTS"IEEE ICACCI, 2015.

[32] Hirshberg,Wang, "Predicting Intonational Phrasing fromText",Proceedings of the 29th annual meeting on Associationfor Computational Linguistics,1991

Normalization of Zhangzhou Citation Tones

Yishan Huang[1], *Mark Donohue*[1], *Phil Rose*[2], *Paul Sidwell*[1]

[1]ANU Linguistics
[2]ANU Emeritus Faculty

yishan.huang@anu.edu.au; mark.donohue@anu.edu.au; philjohn.rose@gmail.com; paulsidwell@gmail.com

Abstract

The seven citation tones of the Southern Min dialect of Zhangzhou 漳州 are described impressionistically, and a linguistic-tonetic representation of their acoustics derived from the z-score normalization of the tones of 9 male and 12 female speakers. A normalization of the raw mean tonal data is shown to be slightly superior to a $\log_{10}$ transform, delivering about an eight-fold reduction in the between-speaker tonal variance (normalization index = 8.6). The data are then used to do a preliminary *Monte Carlo* investigation on how the normalization index changes with the number of speakers used to normalize.

Index Terms: normalization, tonal F0, tonal duration, Zhangzhou dialect.

1. Introduction: Normalization

Fundamental frequency (F0) as the main acoustic correlate of perceived pitch for linguistic systems of intonation, tone and stress shows high variability attributed to a wide variety of factors, both linguistic and non-linguistic [1]. Linguistically, F0 may be intrinsically perturbed by tautosyllabic segments, but also may be perturbed by other contextual factors, for instance the F0 height of segments, or the tone on adjacent syllables, the presence or absence of stress and dialectal accent [1]. Moreover, the F0 on one tone may alternate with another in a specific morpho-syntactic or phonological environment [2]. The non-linguistic factor for F0 variability is well known to be speaker-specific effects [1]. Anatomical and physiological differences in the individual vocal tract structure may generate dynamic acoustic outputs of difference even for phonologically identical utterances. For instance, female speakers generally have higher F0 values than males from their shorter and less massive vocal cords, thus, it is both theoretically and empirically possible for a female's phonological low tone to have a higher F0 value than a male's phonological high tone [3, pp. 38-41].

Given the acoustic variability resulting from differences of individual physiology, it becomes necessary to perform an effective reduction in the between-speaker variance prior to identification of different linguistic categories in signals. The mathematical analog of this process is just what the theory of normalization concentrates on. The aim of normalization is to abstract the variable Individual content away from invariable Linguistic and Accentual content in speech signals, and thereby derive a quantified representation of the variety in question [1].

The main aim of this paper is to present multi-speaker tonal data on the citation tones of the Chinese dialect of Zhangzhou, which has hitherto been impressionistically described a lot, but has arguably received inadequate attention acoustically.

How many speakers are needed for such a quantified investigation of a particular variety of speech? Numbers have been suggested e.g. [4], but not justified. Therefore a subsidiary aim of the paper will be to use the Zhangzhou data to do a preliminary investigation of how the efficiency of the normalization varies with the number of speakers used.

2. Zhangzhou Tones

2.1. Zhangzhou

Zhangzhou is a prefecture-level city situated in mainland China's southern Fujian province with approximately 4.8 million inhabitants [5]. The variety that native Zhangzhou people speak belongs to the Southern Min dialects of the Sinitic language branch of the Sino-Tibetan language family, which are primarily spoken in Southern Fujian and Taiwan.

2.2. Phonetic descriptions of Zhangzhou tones

There have been quite a few previous descriptions of Zhangzhou citation tones [6, 7, 8, 9, 10, 11, 12, 13, 14, 15]. All except two – [6, 10] – agree in describing seven different citation tones. Table 1 lists the descriptions under their Middle Chinese (MC) tonal categories (Ia, IIIb etc. - *a* and *b* stand for historical Yin and Yang registers, respectively). Tonal pitch values are given in the Chao five-point "tone letters" [16]. It can be seen that Zhangzhou does not distinguish separate reflexes of MC IIa and IIb tones (that is the reason only a *II* category is shown), but it otherwise preserves the other six MC categories.

Author	Year	Tone1(Ia)	Tone2(Ib)	Tone3(II)	Tone4 (IIIa)	Tone5(IIIb)	Tone6(IVa)	Tone7(IVb)
Dong	1959	24	212	53	32	33	32	13
Lin	1992	44	13	53	21	22	32	12
Ma	1994	44	12	53	21	22	32	121
FCCEC	1998	44	13	53	21	22	32	121
ZCCEC	1999	44	13	53	21	22	32	121
Gao	1999	45	23	53	21	33	21	121
Zhou	2006	44	13	53	21	22	32	121
Chen	2007	44	13	53	21	22	32	121
Yang	2008	44	13	53	21	22	32	121
Guo	2014	44	13	53	21	22	31	121

Table 1. *Previous descriptions of Zhangzhou citation tones according to Middle Chinese tonal category (in brackets).*

As is usual with impressionistic descriptions there are areas of both agreement and disagreement. Tone 3 for example is uniformly described with a high falling [53] pitch. Disagreements on the pitch values of other tones include, for instance: tone 1 is represented as a low rise [24], a high level [44] or a high rise [45]. Tone 2 is either a low dipping [212] or a low rise [13]. Tone 7 is transcribed as a low rise [13] or a low convex [121] by most scholars. It is not clear what this sort of variation is due to. It could possibly be ascribed to sub-dialectal variation and/or between-speaker or between-transcriber differences.

The previous descriptions are all impressionistic, but in the most recent decade, two scholars [5, 15] investigated the acoustic properties of Zhangzhou citation tones in terms of F0.

However, there are some inadequate and problematic aspects with regard to their research designs and analyses. For instance, [5] compared data from one male and one female directly in terms of their raw F0 values, ignoring the speaker-dependent effects on the F0 realizations and the importance of normalization for linguistic quantitative studies of speech signals. Normalised acoustic data from four speakers were given in [15] using the *T* algorithm, but the study only addressed sonorant-ending tones while neglecting the tones ending in obstruents, and the variation in tonal durations.

In 2015, as part of her Ph.D, the first author, who is a native speaker of Zhangzhou, collected extensive data in the field from 21 Zhangzhou speakers, including, of course, citation tones. It was clear that there were seven citation tones, but the auditory characteristics of most turned out to be different and more complicated from those available in the literature. In addition to pitch differences, the seven tones are characterized by a variety of co-occurring auditory features including length, vowel quality, voice quality, loudness and manner of articulation of syllable-initial consonants. For this paper, however, we will concentrate on their pitch, which it will now be helpful to generalise as follows:

- /**mid rising**/: rising pitch from the speaker’s middle range to high, rather than a high level pitch or a high rising contour as previously described, e.g. /kɔ/ “mushroom 菇”, /si/ “poetry 诗”, /tɐŋ/ “east 东”, /tsɛ̃/ “to contend 争”, /ts^{h}jɐ/ “vehicle 车”, /swẽ/ “mountain 山”, /tĩ/ “sweet 甜”.
- /**low level**/: level in the speaker’s lower third pitch range with long duration, rather than a low rising contour as described before, e.g. /kɔ/ “paster 糊”, /si/ “time 时”, /tɐŋ/ “copper 铜”, /pɛ̃/ “flat 平”, /ɗɐm/ “male 男”, /ɣu/ “cow 牛”, /t^{h}ɐw/ “head 头”, /ts^{h}ɐ/ “wood 柴”.
- /**high falling**/: pitch falling from high in the speaker’s range to low, with a short initial level component. This is similar to previous descriptions but with a lower offset, e.g. /kɔ/ “drum 鼓”, /si/ “to die 死”, /tɐŋ/ “to wait 等”, /ɓɛ/ “horse 马”, /tsjɐw/ “bird 鸟”, /ts^{h}jõ/ “to rob 抢”, /hɐj/ “sea 海”, /tsu/ “host 主”.
- /**mid falling**/: falling from the middle third of the speaker’s pitch range to low, rather than the low falling contour of previous descriptions, e.g. /kɔ/ “to look after 顾”, /si/ “four 四”, /tɐŋ/ “frozen 冻”, /k^{h}o/ “course 课”, /k^{h}wẽ/ “to watch 看”, /hi/ “drama 戏”, /k^{h}ɛ/ “guest 客”.
- /**mid level**/: long and level in the middle third of the speaker’s pitch, rather than at a low pitch range, e.g. /hɔ/ “rain 雨”, /si/ “yes 是”, /tɐŋ/ “heavy 重”, /tjan/ “electricity 电”, /pɛ̃/ “illness 病”, /zi/ “character 字”.
- /**short stopped mid fall**/: mid falling pitch as in the mid falling tone, but with salient short duration, similar to previous descriptions; high vowels are diphthongized, and become creaky to most speakers, e.g. /kɔk/ “country 国”, /sit/ “colour 色”, /kut/ “bone 骨”, /kip/ “urgent 急”, /hwɐt/ “law 法”, /ik/ “one 一”, /ts^{h}it/ “seven 七”.
- /**stopped low level**/: similar pitch to the low level tone, but with a slight final fall due to the depressing effect by creaky phonation. Some rhymes lose their obstruent coda and become open with modal phonation. High vowels are diphthongized. This differs from the low convex contour as described previously, e.g. /tɔk/ “poison 毒”, /sit/ “cooked 熟”, /ɗɐk/ “six 六”, /tit/ “straight 直”, /tsɐp/ “ten 十”, /zit/ “sun 日”, /ɓɔk/ “wood 木”.

It will be noted that the two stopped tones are explicitly treated as separate tonemes, as is usually done in Chinese tonology. The results of the normalization will present acoustic evidence to suggest that they might also be considered allotonically related to two unstopped tones.

3. Procedure

3.1. Speakers and elicitation

The speech data used in this study was collected from a linguistic field trip by the first author in Zhangzhou urban areas of Longwen and Xiangcheng during April 15th to May 9th in 2015. Data was collected from 21 native speakers: 9 males and 12 females. Their ages ranged from 38 to 65, with an average of 54 for males and 53 for females at the time of recording. None had physical difficulties in producing or perceiving speech sounds, and also no difficulties recognizing the words to be elicited.

The recording session was conducted in an acoustically absorbent room of the Zhangzhou Hotel with very little background noise and echo. The words to be read out were shown in simplified Chinese characters by means of *PowerPoint* with one slide per word. One major advantage of using this method is to make sure speakers produce the words in a clear and not-exaggerated voice with balanced and well-controlled intensity and speech rate.

All recordings were digitized at a sampling frequency of 44100 Hz in *Praat* [17] using a professional recorder (*Bluebaby* brand) kindly provided by Huaqiao University. This to a large extent ensured high quality recordings for further linguistic-phonetic processing and analysis.

3.2. Acoustic measurements

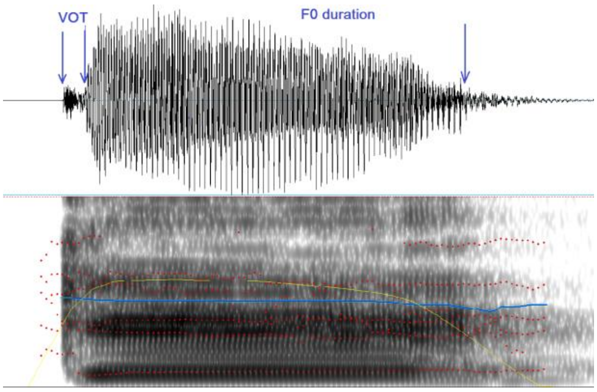


Figure 1: *A Praat labelled example* [ke̞44] “low” *by male Zhangzhou speaker WYF.*

Acoustic measurements were made with *Praat*. The rhyme portion of each monosyllabic token was considered as the tonally relevant F0 duration in this study as shown in Figure 1. The rhyme onset was set at the second strong glottal pulse where waveform amplitude begins to increase and formant patterns for vocalic sounds appear to be stable. The offset was judged to occur at the point where periodicity ceases in the waveform and periodically-excited formant patterns also cease to be visible in the spectrogram.

A *Praat* script was further run to automatically extract tonal duration values and F0 values at 10 equidistant sampling points along the labelled F0 duration. Manual corrections and measurements were made where necessary in the *Praat* pitch window. The extracted raw F0 and duration values were then processed and plotted in *R* for further linguistic-phonetic analyses.

In this study, there were usually 20 tokens for each citation tone. Some tones for some speakers had less than 20 tokens as a consequence of speech errors or unreliably extracted F0 trajectory due to non-modal phonation. This resulted in a total of roughly 29,400 F0 measurements (= 20 examples × 7 tones × 21 speakers ×10 sampling points). Some of the tokens were exemplified in section 2.2 for reference. The tokens were chosen to include as many (sub-)minimal pairs as the phonotactics allowed, and also to maximally reflect phonetic realisations of linguistic tonal categories while balancing the intrinsic perturbation effects on F0 from tautosyllabic segments. Intrinsic vowel F0 was controlled by having comparable numbers of high mid and low vowels.

4. Normalization

Several normalization strategies have been proposed and compared for achieving an effective reduction in the between-speaker variation in tone. For instance, [18] proposed both a z-score normalization on the tones of Standard Vietnamese and a way of evaluating its performance with the normalization index. [19] proposed a T-value transform approach for single speakers of Chinese languages and [20] proposed a revised T-value normalization for a large corpus with multiple speakers. [1] used data from Wu Chinese to argue that z-score normalizations are preferable both on the balance of theoretical considerations and on numerical performance in comparison with fraction of range (FOR) transform strategy. [21] reported the superiority of logarithmic z-score normalization of Shanghai unstopped tones by comparing six different approaches which included z-score transforms, fraction of range (FOR) transforms, proportion of range (POR) transforms, ratio of Logarithmic semitone distances (LD) transforms and logarithmic proportion of range (LPOR) transforms.

As previous studies [1, 22, 25] have demonstrated the superiority of the z-score normalization, this was used on the 21 speakers' raw mean Zhangzhou data, both with and without prior transformation of raw mean F0 values to $\log_{10}$F0. The z-scored normalized F0 values z_i were calculated using the formula:

$$z_i = (x_i\text{-}m)/s \qquad (1)$$

where x_i is an observed F0 value at one sampling point, *m* the mean F0 value and *s* is the standard deviation estimated from all the sampled F0 values of a given speaker's tones.

Normalization parameters of mean and standard deviation were estimated from all sampling points of all tones of all speakers. As proposed in [18], the efficiency of a normalization was quantified with the normalization index (NI), a measure of how well tonal values cluster after normalization. The NI reflects how much the normalization reduces the proportion of variance in a sample due to between-speaker differences in tonal values. The higher the NI value, the greater degree of reduction in between-speaker differences and the clearer the linguistic-phonetic content of the signal.

It was found that the normalization with prior log transform performed slightly worse (NI = 7.7) than the normalization with raw F0 (NI = 8.6). The results for the latter are therefore shown in figure 2, which also plots the normalized F0 values as a function of normalized duration to preserve the relationship between the tonal trajectories [22]. Duration was normalised with reference to the mean duration of all tones [23].

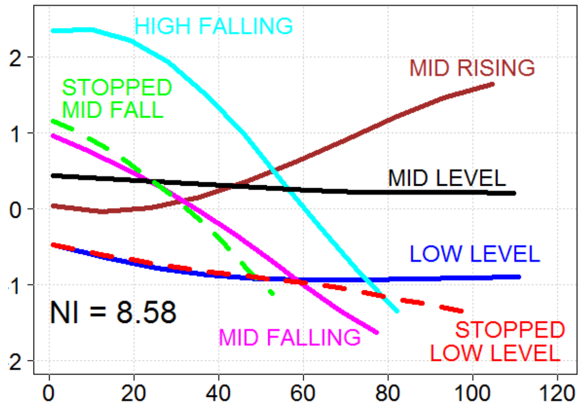


Figure 2: *Intrinsic Z-score normalized F0 for 21 Zhangzhou speakers' citation tones. Y-axis = normalized F0, x-axis = normalized duration (%).*

Figure 2 shows a fairly simple configuration of mean normalized tone trajectories corresponding closely to their impressionistic description. Of interest is the relationship between the stopped and unstopped tonal F0: it can be seen that the F0 of the two stopped tones – stopped mid fall and stopped low level (dashed lines) - is very similar to the F0 of the unstopped mid falling and low level tones respectively. The stopped tones have a slightly shorter duration and a falling offset perturbation, both of which are presumably related to their following voiceless unaspirated coda obstruents or extrinsic laryngealisation. These stopped and unstopped pairs could easily be said to be acoustic allotones of the same tonemes and the configuration could thus be said to comprise two falling, two level and one rising toneme. It is of further interest to note that the pairs of tones with the same contour are not maximally separated. Thus the level tones are not high and low but mid and low; and the falling tones do not fall from high and low but from high and upper-mid. A not too procrustean representation of these shapes in Chao's five point tone letters might be: high falling [51], mid falling and stopped mid falling [41], mid level [33], low level and stopped low level [22], mid rising [35], although this would need to be checked by an appropriate semitone transform, since Chao's tone letters are pitch descriptors and figure 2 is acoustic [24].

5. Optimum speaker number

How many speakers are necessary for achieving an effective estimation of the between-speaker variation and an accurate representation of a particular variety of speech? This is a question that does not seem to have received much attention. For example, in his book *Phonetic Data Analysis*, Ladefoged [4, p. 14] wrote:

> "Ideally you want about half a dozen speakers of each sex. … If you can eventually find 12 or even twenty members of each sex, so much the better."

Ladefoged's figures (24 – 40) were repeated in the *Handbook of Descriptive Linguistic Fieldwork* [p.254]. An ideal database of 30 speakers was recommended in [21] from a statistical perspective, possibly because 30 is about the number where a *t* distribution becomes normal.

In order to investigate how the normalization index changes as a function of the number of speakers, a quasi-*Monte Carlo* approach was used. The F0 trajectories of the tones of the 12 female and 9 male Zhangzhou speakers were first modeled separately with cubic polynomials. The multivariate random command in *R* was then used to generate synthetic sets of 5 unstopped tonal F0 trajectories for 400 male and 400 female speakers from the normal distributions of each polynomial coefficient (multivariate random generation was necessary to take into account any correlation between the coefficients of the tonal trajectories). Random samples of size increasing from 1 male-female pair to 40 were then taken from the 800 speaker data, normalised, and their NI calculated. This was repeated for 30 trials. Figure 3 plots the means and flat-prior credible intervals for the NIs over the thirty trials.

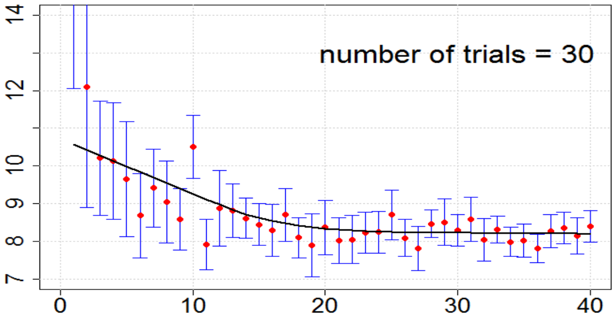


Figure 3: *NI values as function of number of random male-female speaker pairs normalised in 30 trials. Red = means, black = lowess smoothed means, blue = 95% credible intervals. X-axis = number of male-female speaker pairs, y-axis = NI.*

Figure 3 shows that the NI continues to decrease between 1 and ca. 20 speaker pairs (i.e. to 40 speakers), after which the rate of change decelerates. The gradually decreasing credible intervals reflect expected precision increasing with increasing *n*. According to these data, then, the NI is likely to be overestimated in a study with Ladefoged’s lower estimate of six male-female pairs, and his upper estimate of 20 pairs would be needed to give a more accurate estimate. The uncertainties with lower speaker numbers look to be of the order of 3 NI, decreasing to about 1^+ NI by 40 speakers. The effect of normalizing with data unbalanced for sex, as in this paper, remains to be investigated.

6. Summary

This paper has given both impressionistic and acoustic descriptions of the seven Zhangzhou citation tones from 21 speakers. The z-score normalization yielded an eight-fold reduction in the raw between-speaker differences in tonal values in order to extract and specify the Linguistic content of the tonal acoustics. It has also been shown that the two stopped tones have similar enough F0 trajectories to justify an allotonic interpretation of their relationship with the most similar unstopped tones.

7. Acknowledgements

The first author would like to express her thanks to Mark and Paul for their untiring supervisory help and encouragement over the last three years; to Siva for his marvelous help with *R*; and to Phil for the preliminary MC analysis and help in other things *R* and tonal. We also thank our two anonymous reviewers for their time and very helpful comments which we have tried to incorporate.

8. References

[1] Rose, P., "Considerations in the normalisation of the fundamental frequency of linguistic tone," Speech Communication, 6:343-352, 1987.
[2] Chen, M. Tone Sandhi, Cambridge University Press, 2000.
[3] Rose, P. Forensic Speaker Identification, Taylor & Francis, 2002.
[4] Ladefoged, P., Phonetic data analysis: an Introduction to fieldwork and Instrumental techniques, Wiley-Blackwell, 2003.
[5] Yang X., Studies of tones and regional cultures of Zhangzhou dialect 漳州方言声调与地域文化研究, Beijing: Zhongguo Shehui Kexue Chubanshe 中国社会科学出版社, 2008.
[6] Dong T., Four Southern Min varieties 四个闽南方言, Taipei: Zhongyang Yanjiuyuan 中央研究院, 1959.
[7] Lin B., "Zhangzhou vocabularies 漳州方言词汇", Fangyan 方言 ,1-3, 1992.
[8] Ma C., Studies of Zhangzhou dialect 漳州方言研究, Hongkong: Zongheng Chubanshe 纵横出版社, 1994.
[9] FCCEC, Fujian chorography-dialect volume 福建省志-方言志, Beijing: Fangzhi Chubanshe 方志出版社, 1998.
[10] Gao R., "Introduction to the sound system of Zhangzhou 漳州方言音系略说", in Minnan dialect-studies of Zhangzhou variety 闽南方言-漳州话研究, Beijing, Zhongguo Wenlian Chubanshe 中国文联出版社,109-116, 1999.
[11] ZCCEC, Zhangzhou chorography-dialect 漳州市志-方言 49, Beijing: Zhongguo Shehui Kexue Chubanshe 中国社会科学出版社, 1999.
[12] Zhou C.,The great Southern Min dictionary 闽南方言大词典, Fuzhou: Fujian Renmin Chubanshe 福建人民出版社, 2006.
[13] Chen Z., Southern Min dictionary of Zhangzhou variety 闽南漳州腔辞典, Beijing: Zhonghua Shuju 中华书局, 2009.
[14] Guo J., Zhangzhou Southern Min 漳州闽南方言, Zhangzhou: Zhangzhou Library 漳州图书馆, 2014.
[15] Yin X., "Acoustic analysis of tonal patterns in Zhangzhou 漳州话声调格局的分析", Journal of Chifeng University 赤峰学院学报, 30 (6):31-33, 2009.
[16] Chao Y., “A system of tone letters”, Le Maître Phonétique 45: 24-27, 1930.
[17] Boersma, P., “Praat, a system for doing phonetics by computer”, Glot International5:9/10, 341-345, 2001.
[18] Earle, M., An acoustic phonetic study of North Vietnamese tones, Monograph 11, Santa Barbara: Speech Communication Research Laboratories Inc., 1975.
[19] F. Shi, "Tonal studies of disyllaibc words in Tianjin dialect 天津方言双字组声调分析", Studies in languages and linguistics 语言研究, 77-90, 1986.
[20] Shi F., Ran Q., and Wang, P., "On sound pattern 论语音格局" , Nankai Linguistics 南开语言学刊, 1-14, 2010.
[21] Zhu X., "F0 normalization-how to deal with between-speaker tonal variations 基频归一化-如何处理声调的随机差异", Linguistic Sciences 语言科学, 3-19, 2004.
[22] Rose, P., "A linguistic phonetic acoustic analysis of Shanghai tones," Australian Journal of Linguistics, 13:185-219, 1993.
[23] Rose, P., **“**Hong Kong Cantonese Citation Tone Acoustics: A Linguistic-Tonetic Study”, *8th Australian Int’l. Conf. on Speech Science and Technology,* 198-203, 2000.
[24] Rose, P., “Transcribing Tone – A likelihood-based quantitative evaluation of Chao’s tone letters”, Interspeech, Singapore: 101-105, 2014.
[25] Rose, P. “A Comparison of Normalisation Strategies for Citation Tone F0 in Four Chinese Dialects”, Proc. 16th Australasian SST Conf., 2016.

A Comparison of Normalisation Strategies for Citation Tone F0 in Four Chinese Dialects

Phil Rose

Australian National University Emeritus Faculty

philjohn.rose@gmail.com

Abstract

Seven common normalization strategies are compared for unstopped citation tone F0 in the Chinese dialects of Shanghai, Cantonese, Fúzhōu and Zhāngzhōu. A z-score normalization is shown to give clearly superior clustering as quantified by normalization index, but no indication of superiority for a prior log transform of F0 is found.

Index Terms: normalization, tonal F0, tonal duration, Cantonese, Shanghai, Fuzhou, Zhangzhou.

1. Introduction

Quantification, Lord Kelvin famously said, is the first step to science. Speech acoustics, although readily quantifiable, inevitably bear the imprint of the individual vocal tract that produced them, as well, of course, as the various parts of the brain driving that vocal tract. If we are focusing on the speech of the individual, as for example in forensic voice comparison, then this is indeed desirable. But if our focus is Language, then it is often necessary to remove as much speaker-dependent acoustic material as possible so as to arrive at a quantified parametric representation of the variety under question. This is one purpose of normalization: to extract and quantify the Linguistic and Accentual content in the signal by abstracting away from its Individual content. The result should be a quantified representation of the properties of the variety. This is illustrated in figure 1, using data from the 8 male and 8 female Shanghai speakers' unstopped citation tonal F0 in [1].

The left panel of figure 1 plots the 16 speakers' raw mean tonal F0 trajectories as a function of raw mean duration. Apart from the unsurprising fact that the females' tonal F0 generally lies, with a small overlap, higher than the males', the result is rather a mess: it is difficult to see from this figure how many tones there are and what their F0 trajectories are like. The normalisation in the right panel, however, where normalised tonal F0 is plotted against equalised duration, resolves the raw tonal F0 nicely into three groups corresponding to the three unstopped Shanghai tones (often described as high falling, high rising and low rising). Note that the between-speaker differences are reduced, but not eliminated. In particular it appears that there is a sex-related difference in the trajectory of the high-falling tone, which has a much steeper fall in males than females (this may relate to sex-differences in Onset obstruent production and therefore is not necessarily a tonal feature).

A configuration of mean normalised F0 trajectories similar to those in the right panel of figure 1 (shown with thicker black lines) could of course have been obtained without normalization, by simply taking the mean of the raw F0 values in the left panel. However, that would not have allowed an estimate of the variance around the mean normalised curves, which is necessary for many important dialectological, socio-phonetic, typological and even historical purposes, such as quantifying the tones of a variety [1], comparing varieties with respect to their tones [2, 3] or reconstructing tonal acoustics [4]. Neither would it allow for a means of evaluating the efficacy of the normalization, in which variance plays the crucial role.

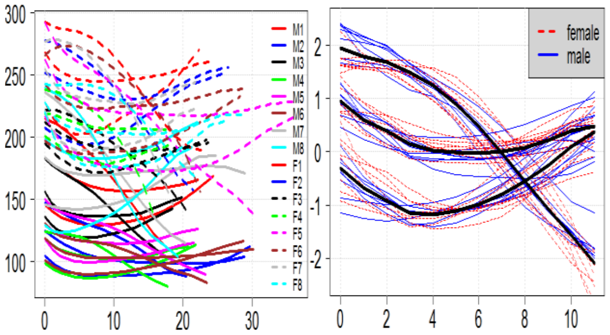


Figure 1: *Normalisation of 16 Shanghai speakers' unstopped tonal F0. Dashed lines = females. Left = mean raw data, right = normalised data. Thicker black lines = mean normalised F0. Axes: left = mean raw duration (csec.) & F0 (Hz.), right = equalised duration (%) & z-score normalised F0 (sds).*

The normalization strategy used in figure 1 is only one of several proposed, but there have not been many attempts to evaluate their performance. The performance of two approaches – z-score and fraction of range – was tested in [5] on Vietnamese, and in [6] on a Yŏngjiāng variety of Wu Chinese. More extensive testing was carried out on Shanghai in [7]. All three studies demonstrated the superiority of a z-score normalization. Even more extensive testing was carried out in [8, p.88ff] on the slope of contour tones from two more Wu dialects, Wúxī and Sōngjiāng. The study, however, was not concerned with overall minimization of between-speaker differences but in determining which normalization minimized sex differences in F0 while preserving age differences *qua* sociolinguistic information. Linear discriminant analysis showed this was best achieved by a simple semitone transform of the tonal F0.

This paper aims to contribute to the evaluation of normalization procedures by seeing which, if any, performs the best on the citation tonal F0 of three more dialects from the major groups of Sinitic: Cantonese, Fuzhou and Zhangzhou, as well as revisiting Shanghai. These data are described in the following section. Section 3 summarizes the main approaches to tonal F0 normalization and section 4 explains the methods for numerically evaluating their performance. Section 5 has the results.

2. Tonal Data

The tonal F0 data used in this paper were taken from previous studies on four varieties from three of the major so-called dialect groups of Chinese: Shanghai (from Wú 吳), Cantonese (Yuè 粵), Fuzhou and Zhangzhou (from Mǐn 閩). Only tones on sonorant-final syllables (舒聲) were used. All the data were controlled to a large extent for intrinsic vowel effects on F0, and are well-balanced for sex.

On syllables ending in a sonorant, conservative Hong Kong **Cantonese** contrasts six tones: three with level pitch, two with rising pitch, and one with falling pitch. The three level-pitched tones are located at the top, in the middle and just below the middle of the speaker's pitch range. Both rising tones start low in the pitch range, with one rising to high and one to mid. The falling tone starts low and falls still lower, such that its phonation type usually becomes non modal as it falls below the speaker's normal pitch range. The data were taken from a linguistic-tonetic description [9] of five female and five male young students recorded in the late nineties. Each tone had 24 tokens balanced for vowel height.

On syllables ending in a sonorant, **Shanghai** contrasts just three tones: one with pitch falling from high in the speaker's range to low; one with dipping pitch in the speaker's mid-pitch range; and one with pitch rising from low in the speaker's pitch range to mid, with or without an initial delay. The data were taken from a linguistic-tonetic description [1] from eight female and eight male students recorded in the late nineties. Each tone had 16 tokens balanced for vowel height.

On syllables ending in a sonorant, conservative **Fuzhou** 福州 is usually described as contrasting five tones [10, pp. 8-9]. There is consensus on the pitch of three: one with pitch falling from high in the speakers' range to low; one with convex pitch in the lower half of the speaker's range and one with level pitch high in the speaker's range. The two remaining tones are in the lower half of the pitch and are variously described as level, falling or rising. In the variety used here, both tones have slightly falling pitch, one in the mid and one in the low range. The Fuzhou data are from five males and five females. Although taken from two studies separated by about 20 years, ([10,11]), the speakers are of comparable age, having been born in the early sixties. The first study had 18 tokens per tone divided equally between [i~ei], [u] and [a] vowels. The second had 3 tokens per tone, all on [a] vowels.

The **Zhangzhou** 漳州 data are taken from a recent study [12] using 12 females and 9 males which shows a five-way tonal contrast on sonorant-final syllables: high and mid falling, mid and low level, and mid rising. Each tone had ca. 20 tokens reasonably well balanced between high, mid and low vocalic nuclei.

Table 1. *Normalisation strategies tested*

strategy	scale	
z-score	linear	$\log_{10}$
FoR_T *max-min*	linear	$\log_{10}$
$\mathrm{FoR}_{T\,PT}$	linear	$\log_{10}$
ST_{meanF0}	semitone	

2.1. Preprocessing

The various sources of the data had used different strategies to sample tonal F0, so all F0 data were pre-processed by first removing any obvious offset perturbations, and then modeling their trajectories with 5th order polynomials. The resulting smoothed trajectories were then resampled at 10% points of duration, as well as at the 5% point, and normalised with different strategies using R code written for the purpose.

3. Typology of normalization strategies

Three basic normalisation strategies can be distinguished for tonal F0. As pointed out in [6], two involve ranges, being of the general form:

$$F0'_i = (F0_i - F0_{\mathrm{ref}})/F0_{\mathrm{range}} \qquad (1)$$

where $F0_i$ is the value to be normalised, $F0_{ref}$ is a shifting factor and $F0_{range}$ is a scaling factor. Of these, a z-score normalization, as its statistical name implies, involves subtracting the value to be normalised $F0_i$ from a sample mean $F0_{mean}$, and dividing by a sample standard deviation $F0_{sd}$, so that the raw F0 values are transformed to multiples of so many standard deviations around a mean of zero:

$$F0.z_i = (F0_i - F0_{mean})/F0_{sd}. \qquad (2)$$

A *z-score* normalisation of tone was first demonstrated on Vietnamese several decades ago [5] and has been used in many studies since. In the second, so-called *Fraction of Range* (FoR) approach, $F0_i$ is expressed as a fraction of the difference between two range-defining points $F0_{upper}$, $F0_{lower}$:

$$F0.For_i = (F0_i - F0_{lower})/(\,F0_{upper} - F0_{lower}) \qquad (3)$$

A version of FoR called *T* [13] is very commonly used in the normalization of Chinese dialect tones, where $F0_{upper}$ and $F0_{lower}$ are a speaker's maximum and minimum F0 values, and the resulting fraction of range value is multiplied by 5 to help mapping onto the well-known Chao five point scale.

The main advantage of z-score normalization is that it ensures a global reduction of between-speaker variation. *FoR*, on the other hand, will force congruence at the range-defining points and thus compromise evaluation of effectiveness in terms of variance reduction [6]. In addition there is the problem of selecting the reference points to use in *FoR*, since it is not known *apriori* which points on a tonal F0 trajectory are comparable between speakers.

The third class of normalization strategy involves a single reference point, usually for semitone transforms. As explained in [8, pp.108-109], reference values can be both fixed, e.g. 100 Hz, or relative to the speaker. Examples of the latter include a speaker's F0 floor, or F0 ceiling, or mid-way between these two; or their overall mean value.

The choice of scale, of course, is logically independent of the normalization approach and can be considered as an additional typological option. A common approach, for example, is to z-score normalise log-transformed F0 values [7]; and the FoR_T normalization also employs a log scale. A final option concerns the source of the normalization parameters: *intrinsic* normalization uses parameters from the data to be normalised; *extrinsic* gets them from elsewhere, for example long-term data [14, 15].

In this paper the three basic approaches were tested: z-score, FoR, and single semitone reference, the first two both with and without prior $\log_{10}$ transforms. Two versions of FoR were tested: a standard FoR_T version, with a speaker's maximum and minimum F0 as range-defining points (FoR_T *max-min)*, and an additional version (FoR_{PT}) with putative comparable pitch target values as more suitable range-defining points. The upper range-defining point was the mean of the high level tone (Cantonese), and the peak of the high falling tone (Shanghai, Fuzhou Zhangzhou); the lower range-defining point was the lowest point of the low to high/ low rising tone (Cantonese, Shanghai), and the mean of the minima in falling tones (Fuzhou, Zhangzhou). The semitone reference method used a speaker's mean tonal F0 as the reference ST$_{meanF0}$ as this was found to be the most suitable in [8]. Table 1 summarises the seven different normalisations tested.

4. Numerical evaluation of normalization

The effectiveness of the normalisation is assessed by the method used in the first tonal normalisation study, on Vietnamese [5, p.133ff.]. Before normalisation the between-speaker variance in raw tonal F0 values will tend to be large because of between-speaker differences in tonal F0 caused by between-speaker differences in mass and length of the vocal folds. A female's high tone may have twice the F0 of a male, for example. After normalisation it is hoped that the between-speaker differences in tonal values will be minimised. Consequently, evaluation of the normalisation strategy involves quantifying how much the normalisation reduces the between-speaker tonal variance in the unnormalised data, a quantity called the *normalisation index* (NI). The idea is to estimate, for both raw and normalised data, the proportion of the overall variance in the data that is due to the between-speaker variance within tones. This is called the *dispersion coefficient*. Since the point of normalisation is to minimise between-speaker differences in tones, the proportion of the overall variance that is due to between-speaker tonal differences is expected to be smaller after normalisation, and so the ratio of the dispersion coefficients for the raw and normalised data – the *normalisation index* – quantifies by how much the between-speaker tonal variance has been reduced and between-speaker differences in tonal F0 have been minimised.

Using the Shanghai data in figure 1 as an example, the calculation of the NI can be formulated thus. Let F0$_{ijk}$ be the F0 value for the i^{th} speaker's j^{th} tone at the k^{th} sampling point. In the Shanghai data for example, i = 1… 16 speakers; j = 1… 3 tones; and k = 1… 12 sampling points (0%, 5%, 10% 20% … 100%). Then the mean F0 value over all speakers at a given sampling point in a given tone $\overline{F0}_{.jk}$ is:

$$\overline{F0}_{.jk} = \frac{1}{16}\sum_{i=1}^{16} F0_{ijk} \qquad (4)$$

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

$$S^2_{\overline{F0}_{.jk}} = \frac{1}{16}\sum_{i=1}^{16}\left(F0_{ijk} - \overline{F0}_{.jk}\right)^2 \qquad (5)$$

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

$$\bar{S}^2_{\overline{F0}_{.jk}} = \frac{1}{36}\sum_{j=1}^{3}\sum_{k=1}^{12} S^2_{\overline{F0}_{.jk}} \qquad (6)$$

For the raw Shanghai data this was ca. 2479. In order to quantify the proportion of the overall variance taken up by variance associated with between-speaker differences in tone, the between-speaker tonal variance is then normalised with respect to the overall variance of the data. This is the mean of the between-speaker variances at each sampling point, i.e. ignoring the tonal differences. For the raw Shanghai data, this was ca. 2804. The ratio of the between-speaker tonal variance to the overall variance is called the *dispersion coefficient* (DC). In this case its value of ca. 2479/2804 = 88% indicates that there is almost as much variation *between* the Shanghai speakers' raw tonal values as in the data overall, and that they effectively do not cluster.

Since normalisation is intended to reduce the between-speaker differences in tonal F0, one expects the DC for the normalised data to be substantially smaller than the DC for the raw data. It is calculated, *mutatis mutandis*, in the same way as the raw DC, namely as the ratio of *between-speaker normalised tonal variance* to *overall normalised variance*. The DC for the normalised Shanghai data was ca. 9%, indicating that only a small amount of the overall variance was taken up by between-speaker differences in tone. The normalisation index (NI) is then defined as the ratio of normalised DC to raw DC. For this normalisation, the NI was 88%/9% = 9, meaning that normalisation has resulted in about a nine-fold reduction in the proportion of variance in the raw data due to between-speaker differences in tone.

5. Results & Discussion

Results are shown in table 2. They agree with previous studies in showing a clear superiority for z-score normalization. Unlike previous results, however, a prior log transform for the z-score is not always preferable: there is nothing to choose between log and linear NIs for Shanghai and Fuzhou; and linear NI is much better than log in Cantonese and a little better in Zhangzhou. This may be because a log transform is over-sensitive to the density of tonal trajectory shapes in the lower pitch range. A FoR$_T$ normalization clearly performs badly if the range is automatically set between a speaker's maximum and minimum F0 values, but can achieve between ca. 50% - 80% of the effectiveness of a z-score with judiciously chosen range-defining pitch targets. The mean semitone transform also performs badly, presumably reflecting large between-speaker differences in their tonal semitone ranges.

Given the large NI difference between Cantonese, with six tones and Shanghai, with three, it is natural to speculate on relationships between NI and number of citation tones or nature of their contrasts. This would be conjecture because NI is a function also of the number of speakers involved and because these are results from single trials: different results would occur purely by chance from further trials with different speakers [12].

Table 2. *Normalisation Indices for strategies tested*

Strategy	Sh	Ca	Fu	Zh
z-score raw	9.8	**21.4**	6.7	**13.0**
z-score log$_{10}$	**9.9**	18.1	**6.9**	12.0
FoR$_T$ *max-min*	4.2	5.9	4.8	7.3
FoR$_T$ *log$_{10}$max-min*	4.2	4.6	4.6	6.0
FoR$_T$ *PT*	7.8	15.9	4.0	7.2
FoR$_T$ *log$_{10}$PT*	7.9	11.9	4.4	6.1
ST*meanF0*	5.8	10.4	3.5	8.0

Figure 2 plots the linear z-score normalised F0 shapes for the four varieties. In anticipation of the paper's final point they are plotted as a function of normalised, not equalised duration (raw duration was transformed to a percentage of a reference duration from the mean duration of all tones [9, 16]). The plots have also been colour-coded to show the different reflexes of Middle Chinese tones in the four dialects.

It looks from figure 2 that, with the possible exception of the high falling tone in Shanghai, Fuzhou and Zhangzhou, and the mid level tone in Cantonese and Zhangzhou, there are no other shared tones between the three varieties. For example, the high level tone appears higher in a speaker's range in Cantonese than in Fuzhou, and the low falling tone in Fuzhou appears to fall at a slower rate than in Cantonese. This may very well be the case, but before one can reliably use such representations to infer linguistic-tonetic (non-) equivalence across dialects one must be sure the normalisation parameters are comparable. Given the different number of tones in each variety and the different way they are distributed, this is unlikely. More research is needed into how normalised representations may be mapped onto each other.

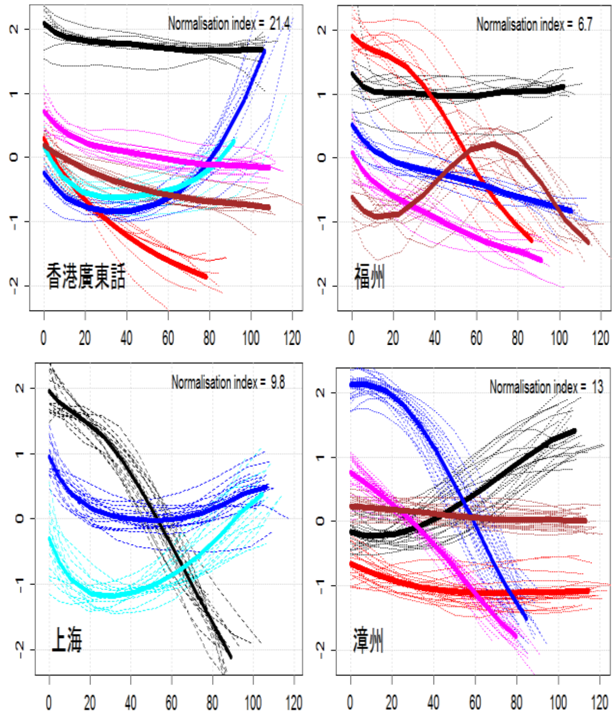


Figure 2: *Linear z-score normalisation of unstopped tones in (clockwise from top left) Cantonese, Fuzhou, Zhangzhou, Shanghai. Middle Chinese tonal reflexes are colour-coded:* **Ia Ib IIa IIb IIIa IIIb**. *Thicker lines = mean normalised F0. X-axes: = normalised duration (%), y-axes = z-score normalised F0 (sds).*

6. Summary & Way Forward

The results of the paper increase the strength of evidence in favour of the superiority of z-score normalization of tonal F0.

The normalisation strategies evaluated in this paper only apply to F0 as a function of *equalised* duration. It may be the case that between-speaker variance in F0 is reduced even more if the normalised F0 is considered a function of *normalised* duration.

As yet, the evaluation of the performance of durationally normalised F0 remains unaddressed. One possible solution is suggested in figure 3, which plots the Shanghai low rising tone F0 as a kernel density distribution. The plot was generated by modelling the distribution of the different speakers' F0 values at every centisecond with a kernel density and then plotting the resulting set of densities as a surface. The tonal density of course increases after normalisation (compare the density axes), and it would be possible to compare the relative amount of increase in durationally normalised and durationally equalised F0.

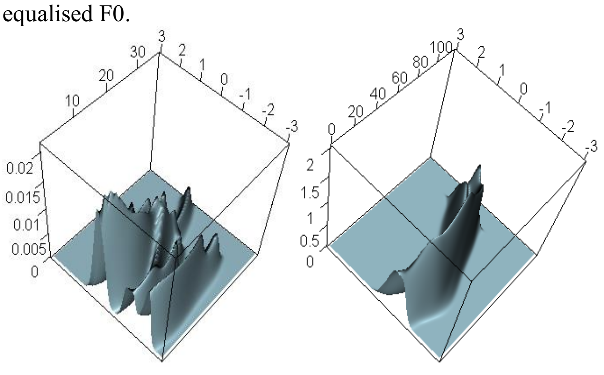


Figure 3: *Density of Shanghai low rising tone F0, left = raw, right = normalised, x-axes = duration, y-axes = z-score normalised F0 (sds), z axes = density.*

7. Acknowledgements

Many thanks to my two anonymous referees for their help!

8. References

[1] Rose, P., "A Linguistic Phonetic Acoustic Analysis of Shanghai Tones", Australian Journal of Linguistics 13:185-219, 1993.

[2] Zhu X., and Rose, P., "Tonal Complexity as a Dialectal Feature: 25 Different Citation Tones from Four Zhejiang Wu Dialects", 5th Int'l Conf. on Spoken Language Processing 3:919-922, 1998.

[3] Steed, W. and Rose, P., "Same tone, different category: linguistic-tonetic variation in the areal tonal acoustics of Chu-qu Wu", Interspeech: 2295-2298, 2009.

[4] Rose, P., "Oujiang Wu tones and Acoustic Reconstruction", in Bowern, Evans, Miceli [Eds.] Morphology and Language History, 235-250, John Benjamins, 2008.

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

[6] Rose, P., "Considerations in the normalisation of the fundamental frequency of linguistic tone", Speech Communication, 6(4):343-352, 1987.

[7] Zhu X. 朱晓农, "基频归一化 – 如何处理声调的随机差异" F0 normalization – how to deal with Between speaker Tonal Variations, 语言科学 [Linguistic Sciences] 3(2):3-19, 2004.

[8] Zhang J., A Sociophonetic Study on Tonal Variation of the Wúxī and Shànghǎi Dialects, LOT Netherlands Graduate School of Linguistics, 2014.

[9] Rose, P., "Hong Kong Cantonese Citation Tone Acoustics: A Linguistic-Tonetic Study", Proc. 8th Australian SST Conf., 198-203, 2000.

[10] Donohue, C., Fuzhou tonal acoustics and tonology, Lincom, 2013.

[11] Peng G., A Phonetic Study of Fuzhou Chinese, Ph.D. thesis, City University of Hong Kong, 2011.

[12] Huang Y., Donohue, M., Rose, P., Sidwell, S. "Normalisation of Zhangzhou tones", Proc. 16th Australasian SST Conf., 2016.

[13] Shi F. 石锋, "天津方言双字祖声调分析" [An analysis of tone in Tianjin disyllables], 语言研究 [Yuyanyanjiu] 1, 1986.

[14] Rose, P., "How effective are long term mean and standard deviation as normalisation parameters for tonal fundamental frequency?", Speech Communication 10:229-247, 1991.

[15] Rose, P., "Mr. White Goes to Market - Running Speech and Citation Tones in a Southern Thai Dialectal", Proc. 15th Australasian SST Conf., 110-113, 2014.

[16] Zhu X., Shanghai Tonetics, Lincom Studies in Asian Linguistics 32, Lincom, 1999.

Perception of tonal contrasts: high-variability perceptual training with iconic orthographic representations

Yan Chen
University of Arizona
yanchen@email.arizona.edu

Abstract

This study examines the effect of orthographic representations for tones on the perception of five Cantonese tone pairs with high perceptual similarities. Native speakers of American English, Mandarin Chinese, and Standard Thai participated in the study. They completed a high-variability AXB pre-test (day 1), a high-variability AX training (day 2 - 4), and a high-variability AXB post-test followed by two generalization tests (day 5). Visual-trained listeners were presented with tone marks resembling the f0 heights and contours of the tones as feedback, whereas non-visual-trained listeners were not. Visual-trained listeners made correct decisions significantly faster in post-test and two generalization tests.

Index Terms: speech perception, perceptual learning, non-native perception, tones, Cantonese

1. Introduction

Laboratory training studies have revealed that perceptual mechanisms can be modified even within a short period of time with laboratory methods, and that perceptual training can direct listeners' attention to previously ignored acoustic features (e.g., [1], [2], [3], [4]). In addition, several studies have shown that a high-variability phonetic training paradigm [5], where listeners are exposed to phonetic variability within a phonetic category (through various talkers), is more effective than a single-talker paradigm and can lead to a more robust category formation. In the domain of tone perception, a high-variability same-different discrimination task improves non-native listeners' perception of Thai mid tones and low tones [6].

The perception of speech can also be influenced by orthography. Diacritic tone marks (e.g. ¯, ´, ˇ, `) have been shown to be facilitatory in processing Mandarin tones. The diacritics lead to significantly higher accuracy in tone production for native English speakers and native Japanese speakers [7] and they facilitate English speakers' learning of pseudo-words with Mandarin tones [8]. However, Mandarin tones all have different f0 shapes and so do their corresponding diacritics, and thus learning the association between Mandarin tones and diacritics may be easy. In order to see whether orthographic representations are truly effective or not, we examine the effect of tone marks on the perception of five tone pairs in Cantonese (a language with 6 lexical tones, shown in Figure 1) with high degree of phonetic similarities:

T2-T5: Both of them are rising tones starting from relatively the same pitch level but rise to different levels. Even some native Cantonese speakers experience difficulties in perceiving these tones [9] and this pair is most often noted as having merged in production [10].

T3-T6: These two level tones differ in f0 height. For a typical voice with f0 covering a range from 140Hz to 275Hz, the difference between T3 and T6 is just about 30Hz. Such a difference is just marginally sufficient to maintain a phonological contrast [11]. Some native speakers tend to merge T3 and T6 in perception and in production [12].

T4-T5, T4-T6, T5-T6: T4, T5, and T6 have very similar starting pitch levels.T6 is a level tone. T4 and T5 only have slight contours toward the end of the syllable. Some native speakers tend to merge T4 and T6 in perception and in production [12].

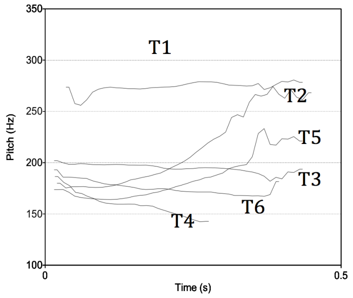


Figure 1: *Real-time Cantonese tones produced by a female native speaker (one token for each tone). The syllable carrying the tones is [si]. T1: High-Level, T2: High-Rising, T3: Mid-Level, T4: Low-Falling, T5: Low-Rising, T6: Low-Level.*

We also examine whether the effect of tone marks is L1-dependent. There languages were chosen because they form a tonality continuum: English (a non-tone language), Mandarin (a tone language with a relatively simple tone system), and Standard Thai (a tone language with five tones, some of which have similar f0 shapes).

2. Methods

2.1. Participants

97 Cantonese-naive participants were recruited as listeners (29 native speakers of American English and 30 native speakers of Mandarin at the University of Arizona, U.S.; 38 native speakers of Standard Thai at Mahidol University, Thailand). None of them had had musical training in the past 6 years (by self-report) at the time of testing. Eight female native speakers of Cantonese were recruited as talkers.

2.2. Stimuli

There were two kinds of stimuli: auditory and visual. For auditory stimuli, 20 Cantonese words were selected: 4 syllables ([fɐn], [si], [jim], [fu]) each carrying each of the five Cantonese tones (T2, T3, T4, T5, and T6). The talkers produced the sentences [ha:22 jɐt5 kɔ:33 t͡si:22 hɐi22] ______ ‘Next character is ______’, where the target words were embedded at the end of the sentences. The target words were excised and normalized for peak amplitude using PRAAT [13]. Words containing the syllables [fɐn] and [si] produced by four of the 8 talkers were used in pre-test, training, and post-test. In generalization test 1 (new syllables with familiar talkers), words with [jim] and [fu] produced by the same four talkers in the training were used. Words with the syllables [jim] and [fu] produced by the other 4 talkers were used in generalization test 2 (new syllables with new talkers). For visual stimuli, five Chao Tone Letters [14] were used: ˧˥ for T2 (High-Rising), ˧ for T3 (Mid-Level), ˨˩ for T4 Low-Falling), ˨˧ for T5 (Low-Rising), and ˨ for T6 (Low-Level). The visual stimuli were used in training as feedback.

2.3. Procedure

2.3.1. Pre-Test

Pre-test was a high-variability AXB discrimination test. For example, a trial testing the contrast of T2 and T5 consisted of the following: [fɐn]2$_{\text{[Talker1]}}$ - [fɐn]2$_{\text{[Talker2]}}$ - [fɐn]5$_{\text{[Talker3]}}$. All possible combinations of the 5 tones were used, resulting in 10 tone pairs. Five tone pairs are considered “difficult pairs” and thus the targets: T2-T5, T3-T6, T4-T5, T4-T6, and T5-T6. The rest were “easy pairs” and they served as fillers. There were 160 trials in pre-test (2 syllable x 10 tone pairs x 8 talker combinations). Inter-stimulus interval (ISI) was set at 1500ms. The listeners were tested individually in a sound-attenuated booth (in US) or a quiet office (in Thailand). The 160 trials were randomly presented over headphones at a comfortable listening level using E-PRIME 2.0 Professional with a desktop computer (in US) or a laptop (in Thailand). The participants were asked to focus on the tones or pitch patterns of the stimuli and make their decisions as quickly as possible. No feedback was given. “A” and “B” responses, as well as reaction times on correct trials were collected. In order to familiarize listeners with the task, a 10-trial practice with feedback was conducted before the experimental trials. The practice section used syllable [jɐu] produced by three female speakers not used as talkers.

2.3.2. Training

The listeners participated in a 3-session perceptual training (1 hour per session, 1 session per day on three consecutive days), starting one day after pre-test. They were randomly assigned to two training paradigms: Auditory-only (AO) and Auditory-Visual (AV): English: AO (n=15), AV (n=15); Mandarin: AO (n=16), AV (n=16); Thai: AO (n=19), AV (n=19). The training made use of a high-variability AX “same-different” discrimination task with the same stimuli used in pre-test. “Same Trials” contained T2-T2, T3-T3, T4-T4, T5-T5, and T6-T6, and “Different Trials” contained only the “difficult pairs” in Pre-test: T2-T5, T3-T6, T4-T5, T4-T6, and T5-T6. For example, a “Same Trial” consisted of [fɐn]2$_{\text{[Talker1]}}$ - [fɐn]2$_{\text{[Talker2]}}$ and a “Different Trial” consisted of [fɐn]2$_{\text{[Talker1]}}$ - [fɐn]5$_{\text{[Talker2]}}$. ISI was set at 1500ms. Feedback was given after the listener responded to a trial. AO listeners saw the correct answer only, while AV listeners saw both the correct answer and the tone marks corresponding to the two auditory stimuli presented on the trial (Figure 2). All listeners were allowed to replay the stimuli as many times as they wanted to. “Same” and “different” responses were collected. There were 480 trials for each training session: 2 syllables x 10 pairs x 12 talker combinations (all possible talker combinations).

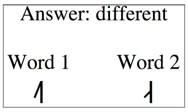


Figure 2: *An example of feedback given to AV listeners. AO listeners received the same feedback except for the tone marks.*

2.3.3. Post-test and Generalization Tests

One day after the last training session, participants took post-test, followed by two generalization tests. The procedure of post-test was identical to pre-test. The generalization tests were high-variability AXB tasks as well. Generalization test 1 consisted of two new syllables [jim] and [fu] produced by the same 4 talkers in training (80 trials: 2 syllables x 10 tone pairs x 4 talker combinations). Generalization Test 2 consisted of the new syllables produced by 4 new talkers (80 trials: 2 syllables x 10 tone pairs x 4 talker combinations).

3. Results

The results from 9 listeners were excluded. Six of them (3 English, 1 Mandarin, and 2 Thai) completed significantly fewer trials on one or more training sessions. Another 2 Thai listeners did not finish post-test due to computer errors.

For each listener at each test (pre-test, post-test, generalization test 1, and generalization test 2), percent correct response was calculated as accuracy. Raw reaction times (RTs) were log-transformed to normalize data distribution. For pre- and post-test performance, percent correct response and log-RTs data each was submitted to a 4-factor mixed-effect ANOVA with L1 (English, Mandarin, Thai) and Training Paradigm (AO, AV) as between-subject factors, and Tone Pair (T2-T5, T3-T6, T4-T5, T4-T6, T5-T6) and Test (pre-test, post-test) as within-subject factors. For performance on the two generalization tests, percent correct response and log-RTs data each was submitted to a 4-factor mixed-effect ANOVA with L1 (English, Mandarin, Thai) and Training Paradigm (AO, AV) as between-subject factors, and Tone Pair (T2-T5, T3-T6, T4-T5, T4-T6, T5-T6) and Test (Gen1, Gen2) as within-subject factors. Results related to the factor Training Paradigm are reported below in details, as it is the main research question in this paper.

3.1. Pre-test vs. Post-test

Figure 3 shows the overall accuracy by L1s, tone pairs, tests, and training paradigms. A 4-way mixed-effect ANOVA showed a main effect of Training Paradigm (F(1,82)=5.38, p<.03), suggesting that AV outperformed AO both before and after training. We concluded that it was a coincidence that AV achieved higher accuracy in pre-test because the participants were randomly assigned to the training groups. Training Paradigm did not participate in any significant interaction. AV did not improve significantly more than AO on accuracy. There were two significant 2-way interactions not related to Training Paradigms: Pair by L1 (F(8,328)=12.99, p<.001) and Pair by Test (F(4, 328)=29.91, P<.001). We found that listeners’ performance on T3-T6 was worse after training than before training (F(1,87)=15.52, p<.001), but there seems to be a trend that AV had higher accuracy than AO in post-test while their performance in pre-test was the same. This suggests that visual training might have prevented listeners from losing more of the level tone distinctions.

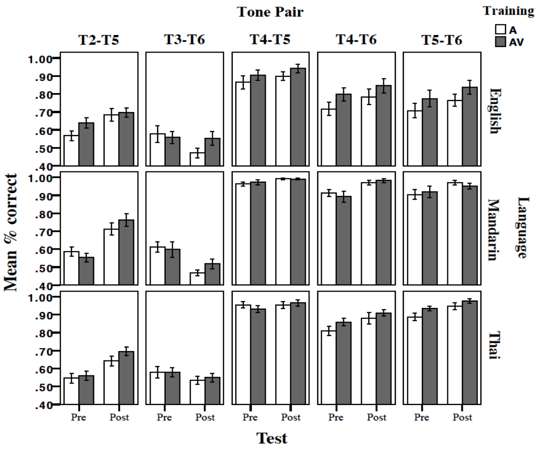


Figure 3: *Mean % correct response by L1s, tone pairs, tests, and training paradigms. Error bars indicate the standard error of the mean.*

For reaction times, Figure 4 shows the overall result of log-RTs by L1s, tone pairs, tests, and training paradigms. The factor Training Paradigm participated in a significant two-way interaction: Test by Training Paradigm (F(1,82)=4.13, p<.05). At pre-test, AV's log-RTs did not significantly differ from AO despite having higher accuracy, as mentioned earlier. At post-test, AV responded significantly faster than AO (F(1,86)=4.63, p<.04) (Figure 5). AV listeners were able to make correct decisions faster than AO listeners after training across the board. The factor L1 participated in two significant two-way interactions, not related to Training Paradigm: Test by L1 (F(2,82)=4.48, p<.02), and Pair by L1 (F(8,328)=9.11, p<.001). The interaction of Test by Pair was also significant (F(4,328)=11.93, p<.001).

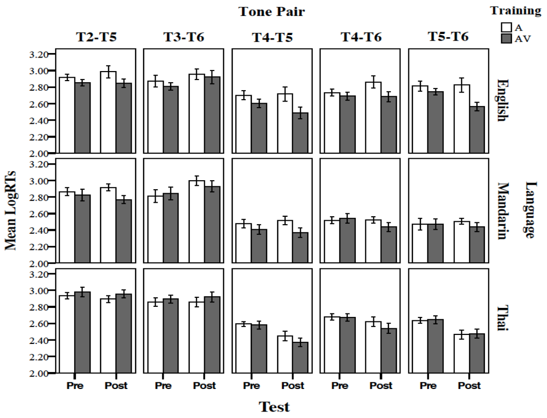


Figure 4: *Mean log-RTs by L1s, tone pairs, tests, and training paradigms. Error bars indicate the standard error of the mean.*

3.2. Generalization Tests

The overall result of accuracy is shown in Figure 6. A 4-way mixed-effect ANOVA did not show a significant main effect of Training Paradigm (F<1), and Training Paradigm did not participate in any significant interaction. This means that AV's per-formance was comparable to AO. The other three factors interacted significantly (F(8,328)=2.84, p<.006).

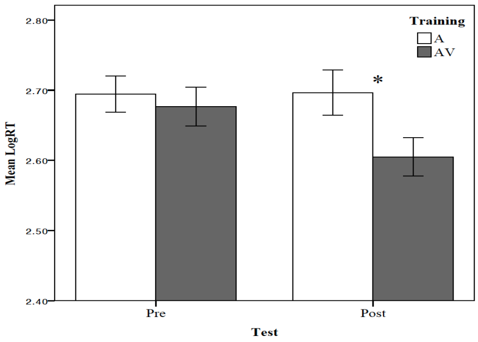


Figure 5: *Mean log-RTs by tests and training paradigms. AV listeners made correct decisions faster than A listeners in post-test. Error bars indicate the standard error of the mean.*

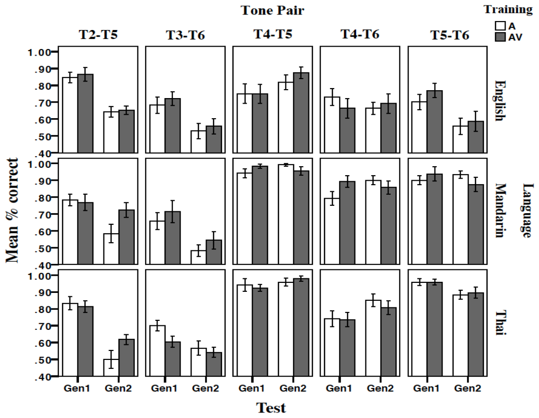


Figure 6: *Mean % correct response by L1s, tone pairs, tests, and training paradigms. Error bars indicate the standard error of the mean. Gen 1 was a generalization test with new words produced by familiar talkers. Gen 2 was a generalization test with new words produced by new talkers.*

For reaction times, the overall result is shown in Figure 7. A significant main effect of Training Paradigm was found (F(1,82)=4.04, p<.05). As shown in Figure 8, AV had significantly shorter log-RTs than AO across L1s, tone pairs, and tests, which means that AV made correct decisions faster when hearing new words from familiar talkers as well as unfamiliar talkers. The lack of Training by L1 interaction suggests that the effect of training was consistent across L1s. In addition to a significant main effect of Training Paradigm, there was a significant two-way interaction of Pair by L1 (F(8,328)=8.67, p<.001).

4. Discussion & Conclusion

The main purpose of this study is to examine whether the use of tone marks in high-variability phonetic training facilitates the perception of tones with small perceptual differences, and whether this kind of orthographic representation is more helpful for listeners from certain L1 backgrounds. Cantonese-naive listeners from three language groups (American English, Mandarin, and Standard Thai) participated in a three-day high-

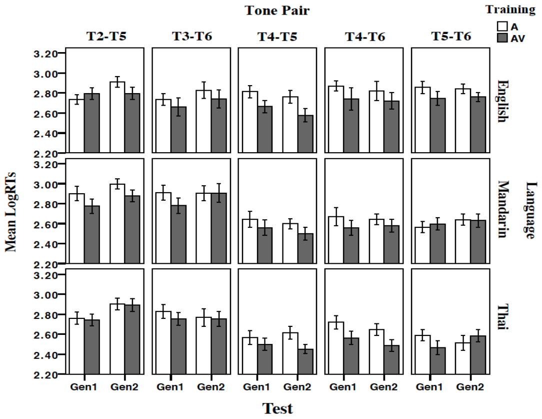


Figure 7: *Mean log-RTs by L1s, tone pairs, tests, and training paradigms. Error bars indicate the standard error of the mean. Gen 1 was a generalization test with new words produced by familiar talkers. Gen 2 was a generalization test with new words produced by new talkers.*

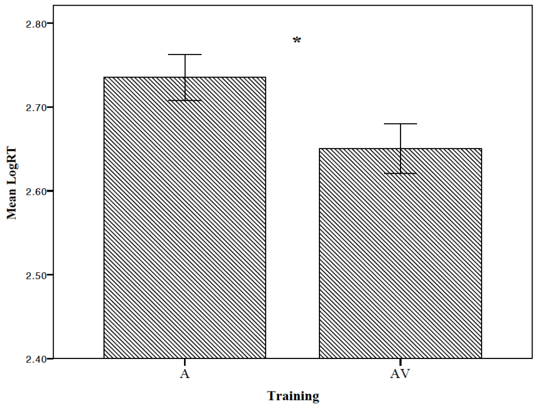


Figure 8: *Mean log-RTs by training paradigms. AV listeners made correct decisions faster than A listeners. Error bars indicate the standard error of the mean.*

variability perceptual training. The fact that training paradigms did not significantly interact with language backgrounds at all suggests that the training had more or less the same effect on both tonal language speakers and non-tonal language speakers.

Both AV and AO listeners improved on accuracy after training, and AV did not have significantly more improvement than AO despite the training with additional information. However, AV's log-RTs on correct trials were significantly shorter than AO in post-test. Tone marks did facilitate the learning of novel tones to some extent, as AV listeners were more certain about the within-category similarities and the between-category differences in the tones and thus could correctly respond faster. The results of the two generalization tests showed similar patterns. AV did not differ from AO on accuracy, but reaction times data showed that not only did AV make correct decisions significantly faster than AO when talkers' voices were familiar (generalization test 1), they also responded faster when talkers' voices were unfamiliar at all (generalization test 2).

The tone marks presented in training are abstractions for F0 contours. Learning to capture the relationship between the auditory stimuli and the visual abstractions may increase listeners' cognitive load. Nevertheless, the mapping between visual abstractions and auditory information helps create better category formation faster. There was a trend that AV had higher accuracy than AO for T3-T6, the only tone pair on which accuracy decreased after training. This indicates that while both groups were impaired by the training on the level tones with high variability, orthographic representations may have prevented AV listeners from further impairment. Although AV listeners formed better tone categories faster, the categories formed were not robust enough for them to outperform AO listeners in terms of accuracy. It should be noted that these results were obtained from 3 hours of high-variability perceptual training (1 hour per day), which is a difficult task in a very short period of time. It is possible that with longer training the facilitatory effect may be shown more clearly.

5. References

[1] Pisoni, D. B., Aslin, R. N., Perey, A. J. and Hennessy, B. L., "Some effects of laboratory training on identification and discrimination of voicing contrasts in stop consonants", Journal of Experimental Psychology: Human Perception and Performance, 8:297-314, 1982.

[2] Logan, J. S., Lively, S. E. and Pisoni, D. B., "Training Japanese listeners to identify English /r/ and /l/: a first report", Journal of the Acoustical Society of America, 89:874-886, 1991.

[3] Wang, Y., Spence, M. M., Jongman, A., and Sereno, J.A., "Training American listeners to perceive Mandarin tones", Journal of the Acoustical Society of America, 106:3649-3658, 1999.

[4] Francis, A. L., Cioccaa V., Ma, L. and Fenn, K., "Perceptual learning of Cantonese lexical tones by tone and non-tone language speakers", Journal of Phonetics, 36:268-294, 2008.

[6] Wayland, R. and Li, B., "Effects of two training procedures in cross-language perception of tones", Journal of Phonetics, 36:250-267, 2008.

[5] Lively, S. E., Logan, J. S., Pisoni, D. B., "Training Japanese listeners to identify English /r/ and /l/ II: The role of phonetic environment and talker variability in learning new perceptual categories", The Journal of the Acoustical Society of America, 94(3):1242-1255, 1993.

[7] McGinnis, S., "Tonal spelling versus diacritics for teaching pronunciation of Mandarin Chinese", Modern Language Journal, 81(2):228-236, 1997.

[8] Showalter, C. E. and Hayes-Harb, R., "Unfamiliar orthographic information and second language word learning: A novel lexicon study", Second Language Research, 29(2):185-200, 2013.

[9] Mok, P. K. and Zuo, D., "The separation between music and speech: Evidence from the perception of Cantonese tones", Journal of Acoustical Society of America, 132(4): 2711-2720, 2012.

[10] Bauer, R. S, Cheung, K. H. and Cheung, P. M., "Variation and merger of the rising tones in Hong Kong Cantonese", Language Variation Change, 15:211-225, 2003.

[11] Hart, J., "Differential sensitivity to pitch distance, particularly in speech", Journal of Acoustical Society of America, 69:811-821, 1981.

[12] Mok, P. K. and Wong, P. W. Y., "Production of the merging tones in Hong Kong Cantonese: Preliminary data on monosyllables", Proceedings of Speech Prosody 2010, Chicago, IL, 2010.

[13] Boersma, P. and Weenink, D., Praat: doing phonetics by computer [Computer program]. Version 6.0.17.

[14] Chao, Y. R., "A system of tone-letters", Le Maitre Phonetique, 30:24-27, 1930.

Exploring the Association of Infant Temperament on Maternal Fundamental Frequency Contours

Alix J. Woolard[1], *Titia Benders*[2], *Linda E. Campbell*[3], *Frini Karayanidis*[1], *Joerg Mattes*[3,4], *Vanessa E. Murphy*[3], *Olivia Whalen*[1], *Alison E. Lane*[3]

[1]University of Newcastle, Australia
[2]ARC Centre of Excellence in Cognition and its Disorders, Macquarie University, Australia
[3]PRC GrowUpWell, University of Newcastle, Australia
[4]Paediatric Respiratory and Sleep Medicine Department, John Hunter Hospital, Australia

{alix.woolard, olivia.whalen}@uon.edu.au, titia.benders@mq.edu.au, {linda.e.campbell, frini.karayanidis, joerg.mattes, vanessa.murphy, alison,lane}@newcastle.edu.au

Abstract

The current study looked at the association between infant temperament and mothers' infant-directed speech regarding adaptations to fundamental frequency (F_o) contours. F_o contours regulate infant attention and affect, and are classified into four contours: rising, bell-shaped, slowly-falling, and rapidly-falling. Eight mother-infant dyads were recruited and participated in a 15-minute play interaction, and mothers' F_o contours were extracted. Infant temperament was assessed using the Temperamental Adjective Triad Assessment. Significant correlations were found between infant approach and rising contours, and infant mood and slowly-falling contours, suggesting evidence of a relationship between infant temperament and mother's F_o contours.

Index Terms: Infant temperament, F_o contours, Infant-directed speech

1. Introduction

Temperament is the differences in reactivity and regulation displayed by an infant, and can include dimensions such as activity, approach to novelty, quality and intensity of mood and attention [1, 2, 3, 4]. An infant's temperamental characteristics are thought to influence the mother-infant interaction, including the mother's linguistic communication [5, 6]. Mothers who speak to their infant automatically use infant-directed speech (IDS), which is a unique speech register characterised by specific semantic and acoustic properties suggested to be fundamental to infant development [7, 8, 9, 10]. It is posited that infant temperament is related to maternal IDS [7, 11, 12]. The current study aimed to look at the association of infant temperament and maternal IDS. Specifically, this study aimed to investigate whether the infant's temperament was associated with an important aspect of the mother's IDS; her fundamental frequency (F_o) contours.

In relation to infant temperament, two functions of IDS are particularly important: attention regulation and affect communication [13, 14]. Attention in infancy generally refers to visual attention, whereby infants demonstrate the ability to track and disengage from a stimulus, orient to a location, and anticipate visual events [15]. Infants display heightened attention when addressed with IDS compared to Adult-Directed Speech (ADS) [9, 16]. Infant attention is also related to affective information, where positive vocalisations elicit infant attention [17].

Affective communication is potentially the most salient function of IDS. Infant affect is expressed through vocalisation, facial expression, and bodily movements, and these expressions normally elicit responses mothers in order to regulate an infant's affect [18]. This affective regulation function in IDS seems to be the most prominent component early on, particularly in the first 12 months [7, 19, 20].

One important difference in IDS compared to ADS involves F_o, which is generally associated with the percept of pitch [21]. F_o is the most salient property of IDS and is predominantly responsible for infant preference for IDS over ADS [11, 22, 23, 24, 25, 26]. Both the attentional and affective function of IDS are largely conveyed by adaptations mothers make using F_o contours [10, 27]. F_o contours typically have one of four shapes when visually represented: rising, bell-shaped, slowly-falling, and rapidly-falling (see Fig.1) [25, 28]. The F_o contour a mother uses corresponds to her communicative intent, which is usually adapted by the mother to bring the infant into an optimal state of arousal and attention [9, 27, 28, 29]. Specifically, rising and bell-shaped contours are associated with positive affect and attaining or maintaining attention, whereas slowly-falling contours are associated with soothing negative affect in infants [25]. Rapidly-falling contours are used as a prohibition against unwanted behaviour.

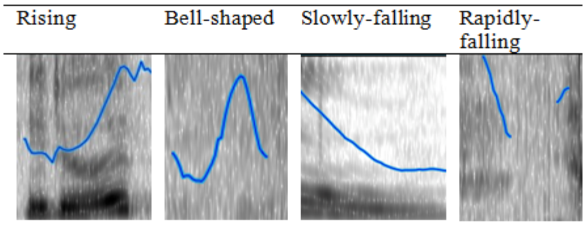


Figure 1: *Typical F_o contours used by the mothers in the current study, the line depicts the F_o contour.*

The occurrences of these four contours have been comprehensively investigated in terms of the mother's side of the communicative interaction [25]. The infant's potential contribution to the mother-infant interaction, however, has not been examined in as much detail. Furthermore, there are few studies that investigate more than one temperamental characteristic in relation to IDS [9, 30], and no previous studies looking at the association between multiple infant temperamental dimensions and the mother's F_o contours.

The current study investigates whether the infant's mood, activity, approach, and intensity are associated with the

F_o contours the mother uses. It is hypothesised that when an infant displays a temperamental trait related to affect or attention, the mother will be more likely to use a F_o contour typically used to communicate affect or regulate attention.

Specifically, we hypothesised that attention-grabbing rising and attention-maintaining bell-shaped contours would be associated with infants who show higher levels of activity and approach. Rising and bell-shaped contours are also assumed to convey positive affect, and as such are expected to be related to high or positive levels of infant mood. Slowly-falling contours that are associated with soothing properties are predicted to be related to a low or negative mood. Rapidly-falling contours, that prohibit unwanted behaviour, are expected to be related to infants displaying a low or negative mood and high levels of activity and approach.

The present study tests a sample of 6-months-old infants, as studies suggest that infants at this age express more emotion and interest compared to older ages [31].

2. Method

2.1 Participants

Infants were recruited through the Breathing for Life Trial-Infant Development (BLT-ID) study, which is investigating the effects of maternal asthma on infant development during the first 12 months of infancy. Eight mothers and their 6-month-old (±30 days) infants (four girls; four boys) participated in the present sub-study, which was undertaken at the Hunter Medical Research Institute (HMRI). The infants did not suffer from any known hearing impairment and mothers consented prior to participation.

2.2 Measures and Equipment

Infant temperament was assessed with the Temperament Adjective Triad Assessment (TATA) [4]. The TATA requires the experimenter to rate the infant on four dimensions of behaviour; Mood, Activity, Approach, and Intensity [4, 32]. These behaviours are represented by 13 items, rated on a Likert scale from 1 to 5, representing opposing polarities. For example, the infant would be rated on the mood subscale from 'Happy/jubilant/cheerful' (1) to 'Sad/Blue/unhappy' (5).

Audio recordings of the mother's speech were made with a Sennheiser ew112 G3 wireless clip-on lapel microphone connected to the mother's clothing to ensure freedom of movement. The microphone sent recordings to an EW100G3 adaptive diversity receiver and a Roland R-26 portable recorder. The speech stream was recorded alongside four video recordings of the interaction captured by Sony HDR-CX405 handycams.

2.3 Procedure

Mothers were informed that the study was investigating how mothers interacted with their children. The mother and infant were seated on a play mat on the floor of the testing room. The mother was given standardised instructions to interact with her infant in as 'natural a way as possible' for 15 minutes. The first 7.5 minutes of the interaction consisted of free-play. After 7.5 minutes, the experimenter placed specific toys chosen for 6 month old infants within the mother's reach for her to interact with her infant in order to facilitate IDS.

2.4 Coding

Two undergraduate students trained in assessment using the TATA rated the infants independently using a scoring template on the four dimensions of temperament; Mood, Activity, Approach, and Intensity. The mean of the raters' scores was used, giving the infant four scores on each of the temperament dimensions.

The audio recordings of the entire interaction were converted into WAV files to be orthographically transcribed and analysed using Praat 5.3.51 [33]. The speech samples were broken down into utterances from the mother, defined as segments of speech separated by more than 300 milliseconds of non-speech [24]. Any utterances that were interrupted by non-speech sounds were omitted from analysis. The F_o contours were extracted from the utterances and graphically rendered in Praat [33]. The coding procedure for classification of F_o contours was adapted from those used in previous studies [28]. A total of 1884 F_o contours were visually classified by a trained researcher into one of the four categories: rising, bell-shaped, slowly-falling, and rapidly-falling. Only two contours were excluded as they were too ambiguous to be classified.

2.5 Design and Analysis

The current study was a cross-sectional exploratory correlational analysis. The data were statistically analysed using the Statistical Package for the Social Sciences [34]. The contours were separated into a 'no toy' and 'toy' condition. We computed per mother, per condition, the proportion of each of the four contours relative to all contours. A Shapiro-Wilk test of normality was conducted on the contour and temperament data sets. Depending on normality, either Pearson's or Spearman's correlations were run on the TATA scores and mothers' F_o contours, to determine if any associations existed between the TATA scores and the proportion of each of the contours.

3. Results

Correlation coefficients for the proportion of each contour class and the infant temperament dimensions are reported in Table 1 for both the 'toy' and 'no toy' conditions.

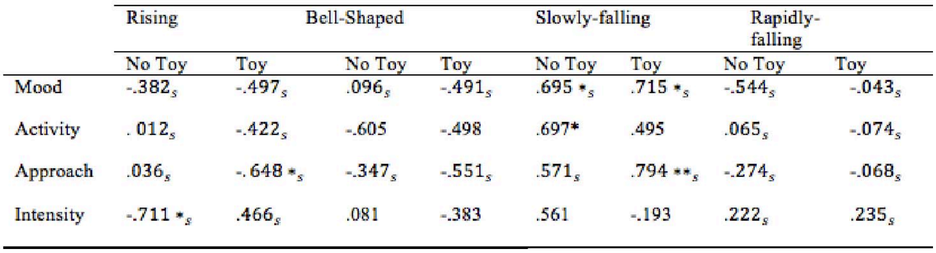
3.1 Rising contours and Infant Temperament

Contrary to the hypothesis, the mothers' proportion of rising contours in either toy condition was not significantly related to either the infants' mood or activity level (see table 1). Although, a trend emerged for fewer rising contours to be used when the infant displayed a more negative mood or less activity. A significant negative correlation was found, however, in the toy condition between the proportion of rising contours and approach (r_s= .690, *N*= 8, *p*= .029; see table 1).

3.2 Bell-shaped contours and Infant Temperament

The mothers' proportion of bell-shaped contours were hypothesised to be associated with infants' mood, activity,

Table 1. *Correlations of Mothers' F_o Contours in the two Toy Conditions and the Infants' TATA Dimension Score*s



	Rising		Bell-Shaped		Slowly-falling		Rapidly-falling	
	No Toy	Toy	No Toy	Toy	No Toy	Toy	No Toy	Toy
Mood	-.382$_s$	-.497$_s$	.096$_s$	-.491$_s$	.695 *$_s$	.715 *$_s$	-.544$_s$	-.043$_s$
Activity	. 012$_s$	-.422$_s$	-.605	-.498	.697*	.495	.065$_s$	-.074$_s$
Approach	.036$_s$	-. 648 *$_s$	-.347$_s$	-.551$_s$	.571$_s$	.794 **$_s$	-.274$_s$	-.068$_s$
Intensity	-.711 *$_s$	.466$_s$	.081	-.383	.561	-.193	.222$_s$	.235$_s$

Note. * indicates a significant correlation at a .05 value, ** indicates a significant correlation at a .01 value, *s* indicates Spearman's correlation, *N*= 8

approach, and intensity. Contrary to predictions, there were no significant relationships found between the mothers' proportion of bell-shaped contours during the two toy conditions and the infants' TATA scores. However, some moderate to strong correlation coefficients did not reach significance (see table 1).

3.3 Slowly-falling contours and Infant Temperament

Slowly-falling contours, used to soothe fussy infants, were expected to be related to infant mood. In line with predictions, mothers did display more slowly-falling contours when their infant was rated as having a more negative mood in the 'toy' condition (r_s= .715, *N*= 8, *p*= .023), and 'no toy' condition (r_s= .695, *N*= 8, *p*= .028; see table 1).

3.4 Rapidly-falling contours and infant temperament

The mothers' proportion of rapidly-falling prohibitive contours were expected to be related to the infant's activity, mood, and approach levels. Contrary to predictions, there were no significant correlations between the proportion of rapidly-falling contours and mood, activity, and approach levels across both toy conditions. Again, some moderate to strong correlation coefficients did not reach significance (see table 1).

4. Discussion

The aim of the current study was to determine if infant temperament was associated with the proportion of rising, bell-shaped, slowly-falling, and rapidly-falling F_o contours that the mother used during a mother-infant interaction.

4.1 Interpretation of Findings

The results of this study provide some support for a relationship between infant temperament and maternal F_o contours. The negative associations between rising contours and the approach dimension of the temperamental assessment may indicate that mothers of infants with a low approach score attempt to increase her infant's attention. These findings are an extension of previous work that suggests rising contours attain infant attention and encourage participation [9, 29]. As low approach has been shown to relate to poorer attentional regulation in infants [35], it follows that infant approach was found to be related to attention-grabbing rising contours.

Infant mood was found to be related to the mother's contours, although differently than predicted. Contrary to predictions, there were no significant associations between rising and bell-shaped contours and infant mood. However, mothers did display more slowly-falling contours when the infants displayed a negative mood during the interaction. Slowly-falling contours are suggested to comfort upset infants, which one can equate with a negative mood. The present results suggest that an infant displaying temperamental characteristics of a negative mood would influence the mother to use F_o contours to counteract negative affectivity [27, 28]. These findings provide a unique insight into the infant dimension of infant-mother communicative interaction, which for the most part has been investigated from the maternal dimension.

4.2 Limitations

We currently interpret the observed associations as reactions by the mothers to their infant's temperament. However, the associations could also indicate that the infants' temperament is a response to the mothers' use of $F_o$contours. We find this alternative interpretation less likely, because F_o contours are suggested to be a functional tool for mothers to use when interacting with their infants [10, 27].

The present study only took one measure of temperament for the entire interaction, and computed the overall proportion of each contour type. However, infant temperament may change over the course of the interaction [4]. These changes may either occur in reaction to the mother's use of contours, or elicit a change in contour use from the mother. A more fine-grained temporal analysis in future research will allow us to draw further conclusions about the direction of influence between mother and infant.

The current study took participants from the BLT-ID study in its early stages, and as such there were issues involving the methodology. One obvious issue was the sample size and subsequent diminished power. In terms of infant studies on F_o contours, recent literature is based on studies with sample sizes anywhere between 10 to 80 participants [14, 19, 24, 31]. Our sample size falls below what is recommended in the published literature. Another issue was that the BLT-ID study is concerned with maternal asthma, thus all mothers who participated were asthmatic. This may have affected their IDS, however due to the time-constraints of the current study we were unable to test a control group. Lastly, there was no inter-rater reliability computed for either the TATA scores or the contour classification. Future studies should take these issues into consideration.

4.3 Significance, Implications and Conclusions

The findings of the current study indicate that infant temperament is linked to maternal IDS during a mother-infant interaction. Investigation into the potential contribution of infant characteristics to IDS is a relatively new subject [28, 31]. Studies that address the infant's influence on the mother can assist in the development of early interventions when the mother-infant relationship is at risk. Children who exhibit a particular temperament may be restricted to hearing a certain type of contour, and thus further research is needed to determine whether this has any long term developmental implications for the infant.

Table of Contents for this manuscript

5. Acknowledgements

We would like to thank all participants, the BLT study and the HMRI. We would also like to thank the following individuals for their assistance on the project: Carly Mallise, Gabrielle Easey, Helen Armstrong and Kelsey Philpott-Robinson.

6. References

[1] Rothbart, M. K., Ahadi, S. A., & Evans, D. E., "Temperament and personality: Origins and Outcomes" in Journal of Personality and Social Psychology, 78:122-135, 2000.

[2] Rothbart, M. K., & Bates, J. E. "Temperament" in Handbook of Child Psychology, Vol. 3, W. Damon, R. M. Lerner & N. Eisenberg, [Ed.], New York: John Wiley & Sons, Inc., 99-166, 2006 .

[3] Carey, W. B. "A simplified method for measuring infant temperament" in Journal of Pediatrics, 77:188-194, 1970.

[4] Seifer, R., Sameroff, A. J., Barrett, L. C., & Krafchuk, E. "Infant temperament measurement by multiple observations and mother report" in Child Development, 65:1478-1490, 1994.

[5] Donovan, W., Leavitt, L., Taylor, N., & Broder, J. "Maternal sensory sensitivity, mother-infant 9-month interaction, infant attachment status: Predictors of mother-toddler interaction at 24 months" in Infant Behaviour and Development, 30:336-352, 2007.

[6] Karrass, J., & Braungart-Rieker, J. M. "Parenting and temperament as interacting agents in early language development" in Parenting, 3:235-259, 2003.

[7] Bornstein, M. H., Tal, J., Rahn, C., Galperin, C. Z., Pecheux, M., Lamour, M., . . . Tamis-LeMonda, C. S. "Functional analysis of the contents of maternal speech to infants of 5 and 13 months in four cultures: Argentina, France, Japan, and the United States" in Developmental Psychology, 28:593-603, 1992.

[8] Burnham, D., Kitamura, C., & Vollmer-Conna, U. "What's new pussycat? On talking to babies and animals" in Neuroscience, 296:1435, 2002.

[9] Fernald, A. "Approval and Disapproval: Infant Responsiveness to Vocal Affect in Familiar and Unfamiliar Languages" in Child Development, 64:657-674, 1993.

[10] Fernald, A., & Mazzie, C. "Prosody and focus in speech to infants and adults" in Developmental Psychology, 27:209-221, 1991.

[11] Fernald, A., & Kuhl, P. K. "Acoustic determinants of infant preference for motherese speech" in Infant Behaviour and Development, 10:279-293, 1987.

[12] Burnham, D., Francis, E., Vollmer-Conna, U., Kitamura, C., Averkiou, V., Olley, A., . . . Paterson, C. "Are you my little pussycat? Acoustic, phonetic and affective qualities of infant- and pet-directed speech" *Paper presented at the Proceedings of the 5th International Conference on Spoken Language Processing: ICSLP'98,* University of New South Wales, Australia, 1998.

[13] Fernald, A. "Intonation and communicative intent in mothers' speech to infants: Is the melody the message?" in Child Development, 60(6):1497-1510, 1989.

[14] Fernald, A. "Meaningful melodies in mothers' speech to infants" in Nonverbal Vocal Behaviour, H. Papousek, U. Jurgens & M. Papousek ,[Eds.], Cambridge: Cambridge University Press, 1992.

[15] Johnson, M. H., Posner, M. I., & Rothbart, M. K. Components of visual orienting in early infancy: Contingency learning, anticipatory looking, and disengaging" in Journal of Cognitive Neuroscience, 3(4):335-344, 1991.

[16] Fernald, A. "Four-month-old infants prefer to listen to motherese" in Infant Behaviour and Development, 8:181-195, 1985.

[17] Singh, L., Morgan, J. L., & Best, C. T. "Infants listening preferences: Baby talk or happy Talk" in Infancy, 3(3):365-394, 2002.

[18] Nicely, P., Tamis-LeMonda, C. S., & Bornstein, M. H. "Mothers' attuned responses to infant affect expressivity promote earlier achievement of language milestones" in Infant Behaviour and Development, 22(4):557-568, 2000.

[19] Benders, T. "Mommy is only happy! Dutch mothers' realisation of speech sounds in infant-directed speech expresses emotion, not didactic intent" in Infant Behaviour and Development, 36:847-862, 2013.

[20] Sherrod, K. B., Crawley, S., Petersen, G., & Bennett, P. "Maternal language to prelinguistic infants: Semantic aspects" in Infant Behaviour and Development, 1:335-345, 1978.

[21] de l'Etoile, S. K. "Infant behavioural responses to infant-directed singing and other maternal interactions" in Infant Behaviour and Development, 29:456-470, 2006.

[22] Soderstrom, M. "Beyond babytalk: Re-evaluating the nature and content of speech input to preverbal infants" in Developmental Review, 27:501-532, 2007.

[23] Ferguson, C. A. "Talking to children: A search for universals" in Universals of Human Language, J. H. Greenberg, C. A. Ferguson & E. A. Moravcsik, [Eds] Stanford, Calif: Stanford Univ. Press, 1:203-224, 1964.

[24] Kitamura, C., Thanavishuth, C., Burnham, D., & Luksaneeyanawin, S. "Universality and specificity in infant-directed speech: Pitch modifications as a function of infant age and sex in a tonal and non-tonal language" in Infant Behaviour and Development, 24:372-392, 2002.

[25] Smith, N. A., & Trainor, L. J. "Infant-Directed speech is modulated by infant feedback" in Infancy, 13(4):410-420, 2008.

[26] Trainor, L. J., & Desjardins, R. N. "Pitch characteristics of infant-directed speech affect infants' ability to discriminate vowels" in Psychometric Bulletin & Review, 9(2): 335-340, 2002.

[27] Papousek, M., Papousek, H., & Symmes, D. "The meanings of melodies in motherese in tone and stress languages" in Infant Behaviour and Development, 14:415-440, 1991.

[28] Katz, G. S., Cohn, J. F., & Moore, C. A. "A combination of vocal F0 dynamic and summary feature discriminates between three pragmatic categories of infant-directed speech" in Child Development, 67(1):205-217, 1996.

[29] Dominey, P. F., & Dodane, C. "Indeterminacy in language acquisition: the role of child directed speech and joint attention" in Journal of Neurolinguistics, 17:121-145, 2004.

[30] Werker, J. F., & McLeod, P. J. "Infant preference for both male and female infant-directed talk: A developmental study of attentional and affective responsiveness" in Canadian Journal of Psychology, 43(2):230-246, 1989.

[31] Kitamura, C., & Burnham, D. "Pitch and communicative intent in mother's speech: Adjustments for age and sex in the first year" in Infancy, 4(1):85-110, 2003.

[32] Seifer, R., Sameroff, A. J., Dickstein, S., Schiller, M., & Hayden, L. C. "Your own children are special: clues to the sources of reporting bias in temperament assessments" in Infant Behaviour and Development, 27:323-341, 2004.

[33] Boersma, P., & Weenick, D. "Praat: doing phonetics by computer. 5.3.51." Retrieved 14 April, 2015.

[34] IBM Corporation. "Statistical Package for the Social Sciences" Armonk, NY: IBM Corporation, 2013.

[35] Carey, W. B., & McDevitt, S. D. "Revision of the infant temperament questionnaire" in Pediatrics, 61:735, 1978.

Emotion-related explanations of the vowel variability in infant-directed speech

Titia Benders

ARC Center for Excellence in Cognition and its Disorders
Department of Linguistics, Macquarie University

titia.benders@mq.edu.au

Abstract

Speech is inherently variable, and so is Infant-Directed Speech (IDS). IDS is also a highly emotional register. In two listener-rating studies of Dutch IDS, we explore emotion-related explanations for (Exp 1) and consequences of (Exp 2) the acoustic differences between IDS vowel tokens. Listeners rated IDS utterances on valence and energy (Exp 1) and on the perception of smiles and child-like speech (Exp 2). The predicted association between valence and formant frequencies was not found (Exp 1), but a higher second formant results in a more smiled and more child-like percept (Exp 2).

Index Terms: infant-directed speech, emotion, affect.

1. Introduction

Parents sound differently speaking to babies than to other adults. Easily perceived are the higher fundamental frequency (F0) and larger F0 range in infant-directed speech (IDS). Possibly less accessible to the naive ear are the acoustic changes to IDS vowels. The corner vowels (/i, u, a/) may be more distinct in IDS than in adult-directed speech (ADS), as measured in an acoustic space defined by the first and second formant (F1, F2) [1]. The corner vowels may also have overall higher formant frequencies in IDS than in ADS [2, 3, 4]. Lastly, the F1 and F2 of individual vowel tokens display larger variation around the mean in IDS than in ADS [1]. The present paper takes a first step towards affective explanations for (Experiment 1) and perceptual consequences of (Experiment 2) the variability in F1 and F2 between individual vowel tokens within IDS.

The acoustic characteristics of IDS can be explained in terms of the three functions that IDS serves: expressing affect, directing attention, and teaching language [5]. The affective function is associated with F0 height and the attentional function with F0 range [6]. Mothers may teach their baby language by enhancing the distances between the corner vowels in IDS and prepare them for between-speaker variability with high vowel variability within their IDS [1].

A different line of research attempts to explain all acoustic properties of IDS as the result of the generally stronger expression of emotion in this register [7]. As IDS primarily expresses a happy emotion, the F0 height and F0 range in IDS may both be consequences of the high energy in very happy speech [8].

IDS as highly emotional speech is difficult to reconcile with enhanced distances between the corner vowels. However, several recent studies have failed to replicate these enhanced distances and report a rise of the formant frequencies in IDS [2, 3, 4]. Raised formant frequencies in IDS could result from the expression of positive affect, as the (positive) facial expression of a smile raises formant frequencies [9]. The raised formants in IDS could also be due to mothers imitating the higher formants in child speech [10]. As adaptation to the interlocutor is a social process, also this latter explanation suggests that the raised formant frequencies in IDS are a consequence of the general expression of positive affect.

No connection has yet been drawn between the expression of emotion and formant variability in IDS. Such a connection would be in line with the idea that the expression of affect impacts on articulation, and thus on formant frequencies [8]. An emotion-related explanation of vowel variability could either replace the language-teaching account of this phenomenon or supplement it: even if vowel variability supports language development, the acoustic properties of each individual vowel token still need to be accounted for.

The present study tests whether the formant variability in IDS vowels is related to the expression of emotion in this register. Experiment 1 tests whether formant frequencies in IDS vowels are higher in utterances with more positive affect. Experiment 2 tests whether IDS utterances with raised vowel formant frequencies are perceived as smiled, child-like, or both.

2. Experiment 1

Experiment 1 tested emotion-related explanations of the IDS vowel formant variability by asking whether formant frequencies in IDS are higher in utterances with more positive affect.

To establish affect in IDS utterances, listeners were asked to rate the Valence and Energy of low-pass filtered IDS and ADS. Valence and Energy scales have been used extensively to study the acoustic correlates of affect [8]. Previous work on IDS has used scales such as Comfort and Direct Attention to understand the acoustic correlates of specific communicative intents [6]. Because the present study aims to understand acoustic variability in IDS in terms of general affective processes, the more general Valence and Energy scales were adopted.

The first sets of analyses established whether perceived Valence and Energy are higher in IDS than ADS, as well as related to the utterance F0 height and range. The final analyses tested whether F1 and F2 are higher in utterances with more positive Valence. The latter analyses only included utterances with the low-back vowel /ɑ/, because IDS causes the largest formant shifts in low [3], or low and back vowels [2, 4]. Predicting formant frequencies from rated affect is not circular, as listeners rated low-pass filtered speech with formants filtered out.

2.1. Methods

2.1.1. Participants

Participants were 15 volunteers from the SONA research participant database at the Radboud University Nijmegen, The Netherlands. All participants were adult monolingually raised native speakers of Dutch, with self-reported normal hearing and (corrected to) normal vision. None of the participants were parents. Participants were compensated with Euro 10 for their participation. Data of 1 participant had to be excluded due to equipment failure. Demographic information from a second participant was missing due to experimenter error, but their data was included the analysis. We thus report data from 14 participants (11 female, 2 male, 1 unknown; 18-25 years old).

2.1.2. Stimuli

The stimuli came from a corpus of 1144 IDS and ADS utterances of 18 mothers who were recorded twice, at their infants' ages of 11 and 15 months [4]. For the IDS recordings, mothers were asked to play with their infants and some provided toys. ADS was elicited by experimenter questions about the infants' familiarity with and interest in the toys. The toy names contained the vowels /i/, /u/, /a/, or /ɑ/ in the stressed syllable (for more details: [4]).

Each utterance was manually re-annotated for its onset, offset, overlapping sounds, and voice quality. The 920 utterances with no sound overlap and a modal voice quality were selected for the present experiment (ADS=109; IDS at 11 months=473; IDS at 15 months = 338).

For presentation in the experiment, the utterances were extracted at zero crossings, low-pass filtered at 400Hz using the pass Hann filter with a smoothing of 100Hz as implemented in Praat [11], and scaled to the same loudness.

For later analyses, we measured the F0 of all utterances and the F1 and F2 of the vowel /ɑ/ as it appeared in the target words. The F0 curve of each (unfiltered) utterance was estimated in hertz using the cross-correlation method. The F0 range for the analysis was initially set at 120–400Hz. If the analysis of the median F0 failed, the F0 floor was updated to 75Hz, and if the analysis still failed, the criterion for the voicedness was lowered from 0.45 to 0.35. From these estimated curves, we computed the median F0 and the F0 range (maximum F0 minus minimum F0). Formants were measured in the central 40% of the vowel using the Burg-algorithm as implemented in Praat [11], which extracted 5 formants per frame with a ceiling of 5577 Hz. The median F1 and F2 were computed.

2.1.3. Procedure

Participants were randomly assigned to one of three experiment versions. Each version contained a third of the stimuli (version 1&2: 307; version 3: 306), equally taken from each speaker's ADS and IDS at both infant ages. Nine stimuli (three from each version) were also presented to all participants as practice trials.

The participants' task was to rate each utterance on two scales. The Valence scale ranged from "fully negative", to "neutral" to "fully positive". The Energy scale ranged from "fully calm", to "neutral", to "fully energetic". The scales were continuous — participants could indicate their rating by clicking anywhere on the scales. Each trial started by playing the stimulus and participants could replay once if desired. The next trial started as soon as the utterance had been rated on both scales.

Participants were instructed that they would listen to mothers speaking to their babies as if "listening through a wall": one cannot understand the words, but still recognise the speaker's affective state. Participants were told that they would rate each utterance for Valence and Energy, and were explicitly reminded that these are separate dimensions of affect.

The experiment started with the 9 practice trials, presented in randomised order. Participants could adjust the sound level during the practice trials and ask any additional questions afterwards. The experiment continued with the 306 or 307 experimental trials, which were randomised for each participant. Participants were required to take 2 breaks, during which they answered some on-paper questions, and filled out an exit questionnaire at the end. The procedure took at most 60 minutes.

The experiment was run in Praat's Demo window [11]

2.1.4. Analysis

The obtained Valence and Energy ratings ranged from 0 (the low end of the scale) to 100 (the high end of the scale). Each utterance was rated for Valence and Energy by 5 or 4 listeners, and these ratings were were averaged to obtain one Valence and one Energy score per utterance.

Data were analysed with linear mixed effects models using *R*'s *lmer* function [12]. Each analysis had one dependent variable and one or more independent variables. Analyses were conducted with different sets of dependent and independent variables, and on various subsets of the utterances. Both the variables and the subsets are addressed in more detail in the Results section. For each analysis, the dependent and the (continuous) independent variables were centred at 0 across all utterances in the subset, irrespective of speaker. All models were fit with by-speaker random intercepts and by-speaker slopes for each independent variable. Covariances between the random effects were not estimated due to convergence problems in some models. Statistical significance of the independent variables was evaluated by treating the t-statistic as a z-statistic, and thus interpreting $t > |1.96|$ as significant at alpha of 0.05.

Table 1: *The scores for Valence and Energy (Experiment 1) and Face and Child (Experiment 2). Mean scores and standard deviations (between parentheses) are computed across all utterances included in the respective analyses.*

experiment	scale	IDS-11	IDS-15	ADS
1	Valence	57.241 (13.428)	57.815 (13.577)	41.421 (13.801)
1	Energy	51.524 (13.815)	53.237 (14.557)	42.52 (16.000)
2	Child	72.604 (13.053)	70.099 (13.749)	20.561 (14.919)
2	Mouth	61.838 (10.11)	60.493 (9.725)	44.747 (10.142)

2.2. Results

The mean Valence and Energy scores for the utterances in ADS, and in IDS to 15- and 11-month olds can be found in Table 1.

The Valence and Energy scores were the dependent variables in two separate analyses conducted on all 920 utterances with the independent variable Register (IDS=0.5 versus ADS=-0.5). Dutch IDS is more positive in Valence and higher in Energy than ADS (Valence: $\beta = 16.425, t = 9.339, p < 0.001$; Energy: $\beta = 10.111, t = 5.296, p < 0.001$). This difference between IDS and ADS in both Valence and Energy validates the scores.

The Valence and Energy scores of the 811 utterances from IDS -thus excluding the ADS utterances- were the dependent variables in two separate analyses with the independent variable Infant Age (11-months=0.5 versus 15-months=-0.5). We found no evidence for differences in Valence or Energy between Dutch to 11- and 15-month-olds (Valence: $\beta = -1.0153, t = -1.062, p = 0.288$; Energy: $\beta = -1.640, t = -1.138, p = 0.255$). The absence of significant Valence and Energy differences between the IDS to 11- and 15-month-olds warrants collapsing these data in the subsequent analyses.

We then regressed the Valence and Energy scores in the 811 IDS utterances on the continuous independent variables F0 median and F0 range. A more positive utterance Valence can be predicted from a higher F0, whereas no association was observed between Valence and F0 range (F0 median: $\beta = 0.03, t = 2.582, p < 0.05$; F0 range: $\beta = 0.009, t = 1.356, p = 0.175$). A higher utterance Energy can be predicted from both a higher F0 and a larger F0 range (F0 median: $\beta = 0.043, t = 4.804, p < 0.001$; F0 range: $\beta = 0.037, t = 4.075, p < 0.001$).

The final, and for the purposes of this study most interesting, two analyses were conducted on the 255 utterances in IDS containing a target word with /ɑ/. The dependent variables were F1 and F2 of the vowel /ɑ/, and the continuous independent variables were the utterance Valence and Energy scores. A higher F1 of /ɑ/ can be predicted from a higher utterance Energy, whereas no association between F1 and utterance Valence was observed (Valence: $\beta = 0.001, t = 0.32, p < 0.749$; Energy: $\beta = 0.011, t = 2.435, p < 0.015$). A higher F2 in /ɑ/ is not reliably associated with either aspect of affect (Valence: $\beta = 0.005, t = 0.96, p < 0.337$; Energy: $\beta = -0.002, t = -0.381, p < 0.703$).

2.3. Conclusion and Discussion

The higher perceived positive valence in Dutch IDS replicates findings on Australian-English [6]. The present results are, to our knowledge, the first to directly show that IDS is perceived as more energetic than ADS. The emotion that is high in positive valence and energy is “happiness”, confirming that IDS may be parsimoniously described as very happy speech [7].

A more positive valence is associated with a higher F0, here and in Australian-English IDS [6]. This casts doubt on the claim that utterance valence is primarily associated with articulation and not F0 [8]. The valence-F0 association may be specific to IDS, or result from a correlation between F0 and an unmeasured valence cue. Alternatively, rating valence in low-pass filtered speech may cause an atypically high reliance on F0.

A higher perceived energy is associated with a higher F0 and larger F0 range, which is in line with the claim that utterance energy primarily affects F0 [8]. Since mothers use both a high F0 and a large F0 range when they encourage their infants’ attention [6], high-energy speech may contribute to the attentional function of IDS.

Contrary to our key predictions, the F1 and F2 variability between vowels is *not* explained by the utterance valence. The absence of the predicted valence-formant association may show that formant frequencies in IDS are *not* higher in utterances with higher positive valence. Specifically in IDS, utterances with positive valence may be produced with spread lips to express happiness or with pouted lips to express comfort. The association between valence and vowel formants might therefore be weaker or more complex than predicted.

Contrary to the theory that that utterance energy primarily affects F0, not articulation [8], we found that F1 variability *is* related to energy. A high F1 is associated with an opened mouth and a surprised open mouth is a frequent facial expression in IDS [13]. Possibly, mothers use high-energy speech in combination with a surprised facial expression to encourage attention. The energy-F1 association may thus arise from a common cause rather than a direct effect of energy on F1.

3. Experiment 2

Experiment 2 assessed the perceptual consequences of the variability in IDS vowel formant frequencies by asking whether utterances with raised vowel formant frequencies are perceived as smiled, child-like, or both.

Listeners were asked to rate the perceived Mouth shape and Child-likeliness of unfiltered IDS and ADS. These new scales were developed for the purposes of the present study. The first sets of analyses established whether perceived Mouth shapes and Child likeliness are higher in IDS than ADS. The critical analyses tested whether utterances with higher formant frequencies are perceived to be produced with a more retracted Mouth shape –and thus with larger smiles– as well as perceived to be more Child like. For reasons described in experiment 1, the latter analyses only included utterances with/ɑ/.

3.1. Methods

Only the differences with Experiment 1 are indicated below.

3.1.1. Participants

Participants in Experiment 2 were 15 volunteers who had not participated in Experiment 1. Data of 1 participant had to be excluded due to equipment failure. We thus report data from 14 participants (10 female, 4 male; 19-30 years old).

3.1.2. Stimuli

Stimuli were the same 920 utterances, but not low-pass filtered.

3.1.3. Procedure

The two scales in this experiment were Mouth and Child. The Mouth scale ran from “pouted lips”, to “neutral mouth”, to “retracted corners of the mouth”. The experimenters provided examples of pouting and smiling lips during their explanation of the Mouth scale. The Child scale ran from “adult-like”, to “intermediate”, to “child-like”.

The instruction did not mention that the speech would sound as if heard through a wall. Participants were asked to base their answer on the tone of voice and not on the content.

3.2. Results

The mean Mouth and Child scores to the utterances from ADS, and IDS to 15- and 11-month olds can be found in Table 1. All analyses reported here have Mouth and Child as the dependent variables in two separate models that are otherwise identical.

Two analyses on all 920 utterances with the independent variable Register (IDS=0.5; ADS=-0.5)] showed that Dutch IDS is perceived to be produced with more retracted lips and in a more child-like manner than ADS (Mouth: $\beta = 16.562, t = 12.276, p < 0.001$); Child: $\beta = 50.624, t = 20.479, p < 0.001$). The difference between IDS and ADS in both Mouth and Child scores validates the scales.

Two analyses on the 811 utterances from IDS with the independent variable Infant Age (11 months=0.5; 15 months=-0.5) showed that IDS to 11-month olds is perceived to be possibly produced with more retracted lips and certainly in a more child-like manner than IDS to 15-month olds (Mouth: $\beta = 1.546, t = 1.872, p < 0.061$; Child: $\beta = 2.866, t = 2.539, p < 0.011$). The (marginally) significant effects of Infant Age warrant maintaining Infant Age as a factor in the subsequent analyses.

The final two analyses were conduced on the 255 utterances from IDS containing a target word with /ɑ/. The continuous independent variables were the F1 and F2 of /ɑ/, the F0 median and F0 range of the utterance, and Infant Age. Each of the acoustic predictors was also allowed to interact with Infant Age. The by-speaker slopes for Infant Age and the interactions were excluded from the models because of convergence problems.

IDS utterances are perceived to be produced with more retracted lips when the vowel /ɑ/ has a higher F2 and when the utterance F0 is higher (F2: $\beta = 1.76, t = 2.797, p- = 0.005$; F0 median: $\beta = 0.026, t = 1.753, p = 0.08$). The Mouth scores were not significantly predicted from either vowel F1 or utterance F0 range (F1: $\beta = 0.221, t = 0.383, p = 0.702$; F0 range: $\beta = 0.014, t = 1.559, p = 0.119$).

IDS utterances are perceived to be produced in a more child-like manner when the vowel /ɑ/ has a higher F2 and when the utterance F0 is higher. (F2: $\beta = 2.923, t = 3.326, p = 0.001$; F0 median: $\beta = 0.044, t = 2.317, p = 0.02$). The Child scores in IDS were not significantly predicted from either vowel F1 or utterance F0 range (F1: $\beta = 0.914, t = 1.022, p < 0.307$; F0 range: $\beta = 0.016, t = 1.575, p = 0.115$). The effects of Infant Age and the interactions between the acoustic predictors and Infant Age were not significant in either model.

3.3. Conclusion and Discussion

These results provide the first evidence that IDS sounds more smiled and more child-like than ADS. IDS is perceived to be more smiled and child-like to 11- than to 15-months-old, which mirrors the higher F2 in IDS to 11-month-olds [4].

Critically, utterances in IDS are perceived as more smiled or more child-like when they contain a vowel with a relatively high F2 or have a higher utterance F0. The present study study thus goes beyond the observation that listeners perceive smiles and frowns from speech with, respectively, a raised and lowered F2 [9], and demonstrates listeners' sensitivity to vowel formant differences in a speech register that is 'happy' across the board.

Interestingly, the aforementioned group effect of infant age on the perception of smiles and child imitation disappears once the vowel F2 and F0 of each individual utterance is considered. The perceived difference between IDS to 11- and 15-month-olds may thus be a direct consequence of the acoustics.

Although smiles and child speech are generally associated with a higher F2 as well as F1 [9, 10], no associations were observed between vowel F1 and the perceptual scores. Note that F1 was not higher in IDS than in ADS in the corpus from which the stimuli were taken [4]. Combined with the present results, this suggests that F2 may be a stronger cue than F1 to smiles and the imitation of child speech, at least in IDS.

4. General Discussion

Two experiments, which constitute the first study on listener ratings of Dutch IDS, explored emotion-related explanations for and perceptual consequences of the acoustic differences between individual vowel tokens within IDS. Experiment 1 failed to confirm the prediction that a more positive utterance valence can explain higher vowel formant frequencies. However, a higher utterance energy can explain a higher vowel first formant (F1). Experiment 2 confirmed the prediction that a consequence of a higher second formant (F2) is the percept of more smiled and more child-like speech. Variability in the production of vowels in IDS is thus associated with some aspects of (perceived) speaker emotion.

These results can inform future studies which directly manipulates speaker affect [14]. Direction manipulation of speaker affect is required to establish that the emotion a speaker wishes to express directly explains her vowel formant frequencies. The present findings would be confirmed if F1 is higher in utterances with high-energy emotions, such as anger or surprise, than in utterances with low-energy emotions, such as sleepiness or calmness, and if F2 is higher when the parent is instructed to smile or imitate their infant. A controlled study will also shed further light on the presence or absence of an association between speaker valence and the vowel formant frequencies in IDS.

5. References

[1] Kuhl, P.K. and Andruski, J.E. and Chistovich, I.A. and Chistovich, L.A. and Kozhevnikova, E.V. and Ruskina, V.L. and Stolyarova, E.I and Sundberg, U. and Lacerda, F., "Cross-Language Analysis of Phonetic Units in Language Addressed to Infants", Science, 277(5044):684–686, 1997.

[2] Englund, K. and Behne, D., "Infant Directed speech in Natural Interaction – Norwegian vowel Quantity and Quality", Journal of Psycholinguistic Research, 34:3: 259–280, 2005.

[3] Green, J.R. and Nip, I.S.B. and Wilson, E.M. and Mefferd, A.S. and Yunusova, Y., "Lip Movement Exaggerations During Infant-Directed Speech", Journal of Speech, Language, and Hearing Research, 53:1529–1542, 2010.

[4] Benders, T., "Mommy is only happy! Dutch mothers' realisation of speech sounds in infant-directed speech expresses emotion,not didactic intent", Infant Behavior and Development, 36:847–862, 2013.

[5] Soderstrom, M., "Beyond Babytalk: Re-evaluating the Nature and Content of Speech Input to Preverbal Infants", Developmental Review, 27:501–532, 2007

[6] Kitamura, C. and Burnham, D., "Pitch and Communicative Intent in Mothers' Speech: Adjustments for Age and Sex in the First Year", Infancy, 4(1):85–100, 2003.

[7] Trainor, L.J. and Austin, C.M. and Desjardins, R.N., "Is Infant-Directed Speech Prosody a Result of the Vocal Expression of Emotion?", Psychological Science, 11(3):188–195, 2000.

[8] Scherer, K.R., "Vocal Affect Expression: A Review and a Model for Future Research, Psychological Bulletin,99:143–165, 1986.

[9] Tartter, V.C. and Braun, D., "Hearing Smiles and Frowns in Normal and Whisper Registers", Journal of the Acoustical Society of America, 96(4):2101–2107, 1994.

[10] Vorperian, Houri K and Kent, Ray D, "Vowel acoustic space development in children: A synthesis of acoustic and anatomic data", Journal of Speech, Language, and Hearing Research, 50(6), 1510–1545, 2007.

[11] Boersma, P. and Weenink, D., "Praat, Doing Phonetics By Computer", [Computer Program], 2014, http://www.praat.org/.

[12] Bates, D. and Maechler, M. and Bolker, B., "lme4: Linear mixed-effects models using S4 classes", [Computer Program], 2012, http://CRAN.R-project.org/package=lme4.

[13] Chong, S.C.F. and Werker, J. and Russell, J.A. and Caroll, J.M., "Three Facial Expressions Mothers Direct to Their Infants", Infant and Child Development, 12:211–232, 2003.

[14] Fernald, A., "Intonation and Communicative Intent in Mothers' Speech to Infants: Is the Melody the Message?", Child Development, 60(6):1497:1510, 1989.

The Role of Affect Processing on Infant Word Learning

Jessica Bazouni[1], *Liquan Liu*[1,3], *Gabrielle Weidemann*[1,2,3], *Paola Escudero*[2,3]

[1]School of Social Sciences and Psychology, Western Sydney University, Australia
[2]MARCS Institute for Brain, Behaviour and Development, Western Sydney University, Australia
[3]Centre of Excellence for the Dynamics of Language, Australian Research Council, Australia

{J.Bazouni, L.Liu, G.Weidemann, Paola.Escudero}@westernsydney.edu.au

Abstract

Although infant-directed speech (IDS) supports infant word learning, it is unclear whether such influence is due to the linguistic or social (i.e., positive emotion) cues conveyed. This paper examines the effect of positive affect on infant word learning. 17-month-old infants were tested on their ability to learn novel word-referent pairings produced in happy adult-directed speech (ADS) via a switch task paradigm. Results showed no evidence of word learning, but there was an increased looking time on test to same and switch trials. The level of affect in ADS may not be sufficient to facilitate word learning.

Index Terms: affect, adult-directed speech, infant word learning, switch task paradigm

1. Introduction

Word learning is a challenging task as infants are required to segment novel words from continuous speech and map them on to a referent (e.g., object) within the environment [1, 2]. When adults communicate with infants they often use simplified words, exaggerated pitch, short utterances, pauses in speech, and positive emotion [3, 4, 5]. This infant-directed speech (IDS) enhances infants' ability to process, segment and learn word-referent pairings [3, 6, 7]. Although positive emotion in IDS is assumed to benefit infant language development, its role in the learning of word-referent pairings is unknown as it is usually confounded with the linguistic properties of IDS [8]. Therefore, it is interesting to examine whether positive emotion facilitates word learning in a non-IDS register, specifically happy adult-directed speech (ADS).

Literature has indicated that the modified acoustic characteristics in IDS, such as hyperarticulated vowels, helps infants recognise clauses in speech, distinguish between different speech sounds, and discern boundaries between words in speech [9 - 11]. For example, research [9] has found that the transitional probabilities in IDS helps 7-month-old infants segment nonsense words from partial words, but not in ADS. This suggests that the linguistic features of IDS provides infants with more information about their native language, which helps them perceive, segment and learn words as opposed to ADS.

However, as affect and infant-directedness co-occur [8], there is a need to investigate the role of positive emotion found in IDS on infant word learning. Researchers [12, 13] have argued that the social nature of word learning, which is comprised of emotional cues, may be equally important as the linguistic properties of IDS. Studies [14, 8] have demonstrated that infants' preference for IDS over ADS appears to be related to the heightened positive emotion. Additionally, research [8] has demonstrated that 6.5-month-old infants showed no preference for IDS compared to ADS when affect was matched across both speech registers. Infants preferred happy speech compared to neutral speech, regardless of the speech register [8]. This suggests that positive emotion may maintain infants' attention and therefore, promote word learning, irrespective of speech register.

So far, researchers [15] have investigated the role of positive emotion within the context of word recognition tasks. For example, a study [15] examined the role of positive affect on word recognition by familiarising 7.5- and 10.5-month-old infants with words presented in happy and neutral affect. During test trials, infants heard familiarised words and unfamiliarised words placed in happy or neutral passages. Results suggest that 7.5-month-old infants attended to the happy passages more than the neutral passages, but were only able to identify the familiar words when affect in test trials were identical to familiarisation. 10.5-month-old infants were able to recognise happy familiarised words but not neutral familiarised words, regardless of test conditions. Additionally, 10.5-month-old infants recognised familiar words when affect was varied across familiarisation and test trials [15]. This suggests that older infants may rely on the affective properties of speech to recognise words across congruent or incongruent passages. Although younger infants were unable to recognise words across unmatched affective passages, the findings nonetheless indicate that emotion, when matched across phases, appears to help younger infants recognise words.

Recently, studies [6, 7] exploring the role of IDS on infant word learning have shown that infants are able to learn novel word-referent pairings presented in IDS, but not in ADS. The age with which IDS continues to support infant word learning has also been investigated. A recent study [7] found that 21-month-old infants learnt novel word-referent pairings better in IDS than in ADS, while 27-month-old infants learnt the pairings in either speech register. Results also showed that 21-month-olds with larger vocabularies learnt the words in either register. This coincides with previous research [2] suggesting that as infants' (18- to 24-months-old) vocabulary increases, they rely less on the properties found in IDS and become more attuned to their native language.

Other researchers [6] have investigated the role of IDS compared to ADS on word learning tasks prior to infants' vocabulary spurt, using a modified Switch Task Paradigm [2]. 17-month-old infants were presented with two same-test trials (i.e., same word-object pairings as in the habituation phase) and two switch-test trials (i.e., violation in word-object pairings). Longer looking time on switch-test trials determined whether infants had learned the word-object pairings. Results indicated that infants learnt the new word-object pairings when presented in IDS but not in ADS [6]. This suggests that features of IDS seem to support infant word learning. However, it is still unclear whether the linguistic or the social properties present in IDS help young infants learn word-referent pairings.

The current study attempts to explore this idea by disentangling the linguistic (i.e., speech register) and social (i.e., positive emotion) nature of IDS. This will also provide insight into the role of positive affect on word learning as its mechanism is unknown. This was explored by asking the question of whether infants could learn novel word-object pairings presented in happy ADS, using the revised switch task paradigm [5]. It was hypothesised that infants would be able to learn the novel word-object pairings presented in happy ADS.

2. Method

2.1. Participants

The final sample size consisted of nineteen 17-month-old infants (mean age 17.34 months, *SD* = .330, range: 16.89 months to 18.20 months; 8 female). Infants were healthy full-term with no family history of dyslexia, or vision, and hearing difficulties. English was the primary language spoken in the household. An additional 14 infants were tested but excluded from the final sample due to fussiness (*n* = 8), inattentiveness (*n* = 2), technical sound issue (*n* = 1), missing data (*n* = 1), and failure to habituate (*n* = 2). Infants were recruited from the MARCS BabyLab database.

2.2. Stimuli and Apparatus

The two isolated words *gabu* and *timay* were produced in happy ADS by a female native English speaker, with prior experience recording IDS and ADS. Thirty-three adult ratings indicated that the words were perceived as happy and spoken in an adult-directed register. Each word consisted of a four-token sequence parted by an 800 millisecond gap of silence. In total, 16 repetitions of each word occurred for a maximum duration of 20 seconds per trial. The labels were used in the habituation and testing phase.

The novel word *lard* [16] was used in the pre- and post-test trial. The word consisted of a 10 token sequence interspersed by a one second gap for a duration of 20 seconds per trial. The pre- and post-test trials were used as a measure of attention. This was explored by examining infants' looking time when presented with a larger acoustic-phonetic change in an auditory label and its associated referent at the start and end of testing session.

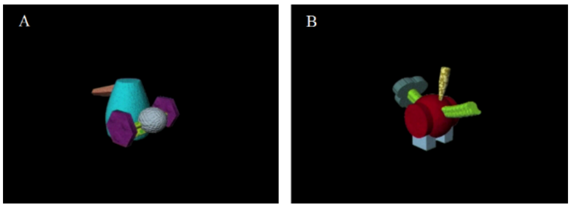


Figure 1. *Image A- Object 1 paired with gabu; Image B- Object 2 paired with timay.*

As shown in Figure 1, the two novel objects [17] paired with the labels *gabu* and *timay* were 3D multi-coloured images, distinct from one another. The word *gabu* was attached to Object 1 (see Figure 1A) and the word *timay* was attached to Object 2 (see Figure 1B). Both novel objects moved back and forth across a black background screen for 20 seconds. The objects' movements were not connected to the spoken words. Both objects were used in the habituation and test phase. A toy waterwheel was presented with the novel word *lard* and was used for the pre- and post-test trials.

Testing was carried out in a small quiet dimly lit laboratory room. A 13.8" monitor was positioned 33.5" in front of the infant. The auditory stimuli were presented at 65 dB ± 5dB amplitude from the infants' seating position. Each testing session was recorded for the purpose of consistent coding and to observe infants' gaze. The experiment was programmed and controlled on the software program Habit X 1.0 on a MacBook Pro laptop. Infants' visual gaze was monitored and recorded using Habit X in an observation room.

2.3. Procedure

Infants were tested separately and sat on their parents' lap. Parents were instructed to not direct the infants' attention back to the screen. As a control, parents were required to listen to music through a cordless headset. In this experiment, a revised version of the switch task paradigm [6] was employed.

Prior to the beginning of each trial, an attention-getting stimulus played to direct the infant's attention to the centre of the screen. Testing session commenced with the pre-test trial, where infants were presented with the novel word *lard* and its associated referent. Following this was the habituation phase during which infants viewed the two word-object pairings separately in a pseudo-randomised order, with no pairing appearing more than twice in a row. This pseudo-randomised order was identical for all infants. The experimenter pressed a key when the infant's gaze was fixated on the screen, and released the key when the infant looked away. The trial continued to play until the infant diverted his or her attention away from the screen for 1 second or a maximum of 20 seconds. The trials ended once the infant had habituated. This was determined by a habituation criterion defined as a 50% reduction in looking time on the last three trials compared to the first three trials. If the infant did not meet this criterion, they viewed a maximum of 25 habituation trials and were excluded from the final sample.

During the test trials, infants were presented with two types of test trials; same-test and switch-test trials. In the same-test trials, infants viewed the same word-object pairings as in the habituation phase. In the switch-test trials, infants were presented with a violation in word-object pairings (e.g., Object 1 paired with *timay*). There were two blocks of test trials which consisted of two same-test trials and two switch-test trials. For half of the infants the initial test block consisted of two same-test trials followed by two switch-test trials. For the other half of infants this order was reversed. This initial test block was then repeated for all infants. The order of the word-pairings in these test trials was counterbalanced across infants. After completing the test trials, infants were exposed to the post-test trial (identical to the pre-test trial).

3. Results

To test the prediction that infants were able to learn word-object pairings in happy ADS, a 2 x 2 (block x test trials) factorial repeated measures ANOVA, with alpha set at .05, was performed on the mean looking time of same-test trials versus switch-test trials. As can be seen in Figure 2, there was no evidence of looking time differences in the first block of test trials compared to the second block of test trials, F(1, 18) = 1.855, p = .190, η_p^2 = .093 . Figure 2 also shows no indication of a difference in looking time on same-test trials compared to switch-test trials, F(1, 18) = 1.347, p = .261, η_p^2 = .070. There was no interaction between blocks of trials and test trials, F(1, 18) = .000, p = 1.000, η_p^2 = < .001. These results indicate an absence of evidence of word learning in happy ADS.

A dependent-samples t test did not reveal a difference in average looking time on the pre-test trial (M = 18.14, SD = 1.53) compared to the post-test trial (M = 18.35, SD = 1.66), t(18) = -.355, p = .726, 95% CI [-1.419, 1.008]. This indicates that infants' level of attention was the same from the beginning to the end of the testing session. The average number of trials to reach the 50% habituation criterion was 10 trials (SD = 4.94), and the average total looking time to reach habituation was 98.35 seconds (SD = 53.22).

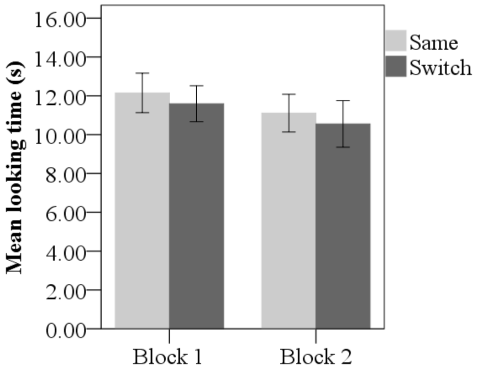


Figure 2. *Mean looking time on same-test trials versus switch-test trials across test blocks. (Error bars represent +/- SEM)*

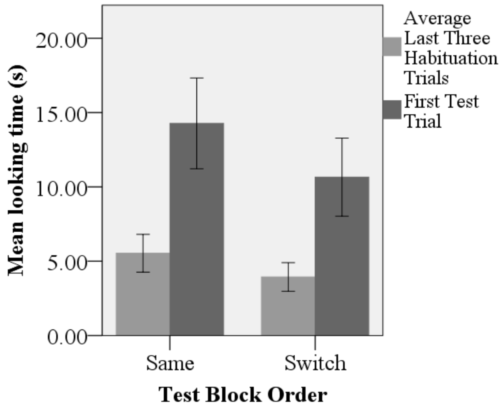


Figure 3. *Looking time on the last three habituation trials (averaged) to the first test trial. (Error bars represent +/- SEM)*

To explore why there may have been no difference between same and switch test trials, a mixed repeated factorial ANOVA, with alpha set at .05, was performed on the average looking time on the last three habituation trials compared to the first test trial, across test blocks. As can be seen in Figure 3, infants showed a significant increase in looking time on the first test trial compared to the last three habituation trials, F(1, 17) = 84.210, p = < .001, η_p^2 = .832. There was no difference between the presentation of test trials (same-test or switch-test) across the test blocks, F(1, 17) = 3.993, p = .062, η_p^2 = .190. An interaction effect was not found between the test phases and order of presentation in test trials, F(1, 17) = 1.438, p = .247, η_p^2 = .078. The results indicate that infants showed renewed interest to the word-referent pairings from the habituation phase to the test phase, regardless of the order of presentation.

4. Discussion

Previous studies [6, 7] evaluating the role of IDS showed that younger infants were able to learn novel word-referent pairings in IDS, but not in ADS. The main objective of this study was to examine the role of positive emotion in the context of learning word-referent pairings by using an ADS register. Against our initial hypothesis, results with 17-month-old infants demonstrate an absence of evidence when learning the word-object pairings presented in happy ADS. The null result was unexpected, as previous work [6, 2] using the switch task paradigm has shown robust effects in infant word learning.

Further inspection of the results suggests that the lack of statistical evidence may be due to a dishabituation effect occurring between the last three habituation trials and the first test trial, regardless of the test block infants were presented with. The reason of this effect is unclear. One possibility is that infants' fast pace of habituation (i.e., brief exposure to the word-referent pairings) might have led to a zone of familiarity preference during the test trials [18]. This might have compromised the results since infants looked longer in the same-test condition. Despite a decrease in looking time in the habituation phase, infants might not have habituated. Rather, infants were drawn to the familiarity of the word-object pairings which extended renewed interest during test phase resulting in sensitised looking time between the same- and switch-test trials. Further investigations are needed to disentangle the effect of the dishabituation.

Alternatively, it may be argued that infants primarily rely on the linguistic and structural properties of IDS to learn new words as it provides more information about the structural patterns of their native language than positive emotion [9, 10, 19]. In agreeance, studies [6, 7] have demonstrated that infants are able to learn novel word-referent pairings when the words are produced in IDS, but not in ADS. However, others [12, 13] have argued that infants not only rely on the linguistic properties, but also the social and emotional cues conveyed in speech to learn new words. Accordingly, prosody not only refers to the linguistic structure of language, but also to the paralinguistic information such as intention and emotion, which convey meaning about language [20].

This may possibly indicate that happy ADS in the present design was not effective under the conditions tested (i.e., infants looked similarly to the same- and switch-test trials). Positive emotion in the current study was not matched to the level of positive emotion found in IDS, or in previous research [e.g., 8]. For example, studies [8] have found that infants do not prefer an infant-directed or adult-directed register when affect is matched. Rather, infants prefer happy speech compared to neutral speech. This indicates that infants might prefer the heightened positive emotion present in IDS [14], as they rely on its saliency to help detect, process and encode certain speech in their native language. This explanation is further strengthened by studies [15] demonstrating that positive emotion helps infants recognise certain words produced in passages. Additionally, studies [21] have shown that adults remember emotionally charged words more than neutral words. While the current study did not show evidence of word learning, the present discussion is noteworthy as it suggests that expressive communication may encourage attentional processing and word learning.

The view that positive affect may encourage word learning is strengthened by studies [22, 23] exploring maternal depression and infant language and cognitive development. Studies [22, 23] investigating IDS produced by mothers with clinical depression suggest that a deficiency in emotional communication may cause deficits in language development, as infants are less likely to employ associative learning strategies. A recent study [22] examined whether there was a correlation between maternal depression and infant language and cognitive development. Results showed that an absence of expressive communication correlated with deficits in infant cognitive development. This suggests that the positive emotion found in IDS appears to play an important role in regulating and encouraging language and cognitive development.

A limitation of this research is that an IDS condition was not employed to compare with the current happy ADS condition. This could have provided possible reasons as to why infants showed no evidence of word learning. Affect could have been matched and compared to an IDS condition to investigate whether positive emotion encourages word-referent learning. Future research should explore this and also test happy ADS to a neutral IDS condition. This could help explain the mechanism of positive emotion and provide additional information on whether infants rely on the linguistic and social properties of IDS mutually or separately to learn words.

5. Summary and conclusion

The study is important to current research on early word learning as minimal research has investigated the role of positive affect supporting infant learning of word-referent pairings. This was explored by disentangling the linguistic and social nature of IDS. An important implication of this study is that IDS, to some extent, appears to maintain certain intrinsic features which separates it from ADS. It is, therefore, interesting to continue exploring the role of positive emotion found in speech on infant word learning, particularly as previous research suggests that positive emotion may be an important feature facilitating language development.

6. Acknowledgements

We gratefully acknowledge the parents and infants who willingly participated in this research. We thank Nicole Traynor, Christina Quattropani, Candice Michael, Scott O'Loughlin, Isabel Lopez and Valeria Peretokina for help with participant recruitment and testing. This study was funded by a grant from the ARC Centre of Excellence for the Dynamics of Language awarded to the third and last author.

7. References

[1] Quine, W. V. O., "Word and object", Cambridge, MA: MIT Press, 1960.

[2] Werker, J. F., Cohen, L. B., Lloyd, V. L., Casasola, M. and Stager, C. L., "Acquistion of word-object associations by 14-month-old infants", Developmental Psychology, 34(6): 1289-1309, 1998.

[3] Saint-Georges, C., Chetouani, M., Cassel, R., Apicella, F., Mahdhaoui, A., Muratori, F., Laznik, M-C. and Cohen, D., "Motherese in interaction: At the cross-road of emotion and cognition?", PLoS ONE, 8: e78103, 2013.

[4] Dunst, C. J., Gorman, E. and Hamby, D. W., "Preference for infant-directed speech in preverbal young children, Center for Early Literacy Learning", 5(1): 1-13, 2012.

[5] Segal, J. and Newman, R. S., "Infant preference for structural and prosodic properties of infant-directed speech in the second year of life", Infancy, 20(3): 339-351, 2015.

[6] Graf-Estes, K. and Hurley, K., "Infant-directed prosody helps infants map sounds to meanings", Infancy, 18(5): 797-824, 2013.

[7] Ma, W., Golinkoff, R. M., Houston, D. M. and Hirsh-Pasek, K., "Word learning in infant- and adult-directed speech", Language Learning and Develop., 7(3): 185-201, 2011.

[8] Singh, L., Morgan, J. L. and Best, C. T., "Infants' listening preferences: Baby talk or happy talk?", Infancy, 3(3): 365-394, 2002.

[9] Thiessen, E. D., Hill, E. A. and Saffran, J. R., "Infant-directed speech facilitates word segmentation", Infancy, 7(1): 53-71, 2005.

[10] Wang, Y., Seidl, A. and Cristia, A., "Acoustic-phonetic differences between infant- and adult-directed speech: the role of stress and utterance position", J. of Child Language, 42(4): 821-842, 2015.

[11] Kuhl, P. K., Tsao F-M. and Liu H-M., "Foreign-language experience in infancy: effects of short-term exposure and social interaction on phonetic learning", Proc. of the Nat. Academy of Sciences, 100(15): 9096-9101, 2003.

[12] Tomasello, M., "The social-pragmatic theory of word learning", Pragmatics, 10(4): 401-413, 2000.

[13] Bloom, L., "Language development and emotional expression", Pediatrics, 102(E1): 1272-1277, 1998.

[14] Panneton, R., Kitamura, C., Mattock, K. and Burnham, D., "Slow speech enhances younger but not older infants' perception of vocal emotion", Research in Human Develop., 3(1): 7-19, 2006.

[15] Singh, L., Morgan, J. L. and White, K. S., "Preference and processing: The role of speech affect in early spoken word recognition", J. Memory and Language, 51(2): 173-189, 2004.

[16] Escudero, P. Best, C. T., Kitamura, C. and Mulak, K. E. "Magnitude of phonetic distinction predicts success at early word learning in native and non-native accents", Frontiers in Psychology, 5: 1059, 2014.

[17] Yildrim, I. and Jacobs, R. A., "Transfer of object category knowledge across visual and haptic modalities: experimental and computational studies", Cognition, 126(2): 135–148, 2013.

[18] Houston-Price, C. and Nakai, S., "Distinguishing novelty and familiarity effects in infants preference procedures", Infant and Child Develop., 13(4): 341-348, 2004.

[19] Soderstrom, M., Blossom, M., Foygel, I. and Morgan, J. L., "Acoustical cues and grammatical units in speech to two preverbal infants", J. of Child Language, 35(4): 869–902, 2008.

[20] Friend, M., "The transition from affective to linguistic meaning", First Language, 21(63): 219-243, 2001.

[21] Kensinger, E. A., & Corkin, S., "Memory enhancement for emotional words: Are emotional words more vividly remembered than neutral words?", Memory and Cognition, 31(8): 1169-1180, 2003.

[22] Kaplan, P. S., Danko, C. M., Everhart, K. D., Diaz, A., Asherin, R. M., Vogeli, J. M. and Fekri S. M., "Maternal depression and expressive communication in one-year-old infants", Infant Behaviour and Develop., 37(3): 398-405, 2014.

[23] Kaplan, P. S., Bachorowski, J., Smoski, M. J. and Hudenko, W. J., "Infants of depressed mothers, although competent learners, fail to learn in response to their own mothers' infant-directed speech", Psychological Sci., 13(3): 268–271, 2002.

Exploring quantitative differences in mothers' and fathers' infant-directed speech to Australian 6-month-olds

Christa Lam-Cassettari & Paige Noble

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

c.lam-cassettari@westernsydney.edu.au

Abstract

Children vary greatly in the rate at which they acquire language in the first years of life (1). A growing body of research indicates that the quantity of parental speech input significantly influences individual differences in child language development (2-7). This study uses the Language Environment Analysis System (LENA) to explore the relationship between the quantity of mothers and fathers speech input and infant language development in a group of Australian infants. Results from 10-14 hour recordings of 11 6-month-old infants reveal that turn-taking quantity is positively related to the quantity of child vocalisations.

Index Terms: infant-directed speech, mothers, fathers, language development, pre-linguistic communication, LENA

1. Introduction

Infants show a remarkable ability to acquire language in their first years of life, often without strict instruction. A growing body of literature indicates that a key factor in this early social environment is the quantity of speech input provided to young children. In fact, speech quantity has long-term influences on their language development and subsequent school readiness (2-7). In the seminal work of Hart and Risley (2), the quantity of parental speech input recorded in hour long recordings was directly related to the child's academic success and intelligence. Furthermore, the quantity of parental speech input was more important than parent education and socioeconomic status in predicting intelligence, vocabulary and language-processing abilities.

Other studies examining the influence of mothers' speech on child language outcomes show that *reciprocal* vocalisations between mother and child positively influence child language (6). Specifically, caregivers who engaged more frequently in reciprocal communication episodes with their children promoted better vocabulary development. More recently, the quantity (amount) and quality (diversity) of parental speech input to 21-month-olds was shown to positively influence child language outcomes at 27 months of age (5).

Many studies examining how factors in the early language environment are related to subsequent language development are conducted with older infants and toddlers to determine how environmental factors influence normed speech and language measures, which generally include vocabulary production. The recent development of the LENA (Language ENvironment Analysis System) makes it easier to examine the speech input and early vocalisations of very young pre-linguistic infants. The LENA is a small recording device that can be worn by an infant or child and record up to 16 hours of audio recordings in the home environment. The recorders are then connected to patented software developed by the LENA Research Foundation to measure the amount of parental speech and other auditory input in a child's natural environment, in addition to the child's vocal output. A recent longitudinal study examining parental speech input to American infants at birth (before leaving the hospital), 44 weeks postmenstrual age, and 7-months used the LENA to evaluate differences in an infant's language environment according to whether the infant was male or female, as well as whether the caregiver was the mother or father. Johnson and colleagues found that mothers provided more overall language input (average word count) than fathers (7). Furthermore, infants showed preferential vocal responses to their mothers in the first months of life (7). Therefore variation in the quantity of speech input is already evident from birth.

With consensus that the quantity and quality of speech input play a crucial role in shaping infant language outcomes, it is necessary to examine precisely how speech quantity and quality affect early language. Focusing on the pre-linguistic period will provide an opportunity to detect where early differences occur so that children can be given ample opportunity to receive speech input that will promote the best start to life. This study explores whether quantitative differences in the infant's language environment accounts for individual differences in the early vocal behaviours of pre-linguistic Australian infant's. The present study addresses two research questions:

1. Is there a difference in the quantity of maternal and paternal language input to pre-linguistic infants?

2. Are individual differences in early infant vocalisations associated with quantitative aspects of parental input including total number of words the infant hears and frequency of turn-taking episodes?

3. It was hypothesized that mothers would produce more speech input than fathers (7), and the quantity of speech would be positively related to the amount of infant vocalisations (2-7).

2. Method

2.1 Participants

The current sample consists of 12 families (mother, father and infant) that were recruited from the MARCS Institute Baby register at Western Sydney University and currently participating in a longitudinal study examining the influence of qualitative and quantitative factors on individual differences in word learning in the first two years of life. All infants were delivered full-term (38-42 weeks gestation), had an average birth-weight of (3.51 kilograms), and had no major birth or postnatal complications. At the time of testing all infants were 6 months of age, and were reported to be healthy with no history of ear infections or any hearing loss. All mothers included in the study had a university degree and had an average age of 33 years. On average, all fathers had graduated high school and completed some type of post school qualification and had an average age of 35 years (see Table 1 for further details). All families spoke English at-home and were monolingual. However, it should be noted that one father had a British accent. Families received a $30 travel reimbursement, a Baby lab degree and an age appropriate gift for the infant in return for their participation in the study.

Table 1. *Demographic information*

Infant Age	6.2 months (range 5.65-6.71 months)
Sex	6 female; 6 male
Birth weight	3.5 kilograms (range 2.9-4.0 kilograms)
Parent age	Mothers (28-39 years) Fathers (27-42 years)
Infant Nap duration	2 hours, 20 mins day (range 1.20-4.20 hours)
LENA duration	Mothers (10-14 hours) Fathers (10-14 hours)

2.2 Materials

The LENA digital language processor (DLP) is a large unobtrusive recording device that is worn by the infant in a specialized vest. The vest enables the child to comfortably wear a small digital audio recorder for up to 16 hours in the child's natural speech environment. After a full day of recording, the LENA DLP is plugged into the LENA computer software program that is designed to automatically segment and analyse the audio recordings and provide quantitative reports on specific measures of the child's language environment. See https://www.lena.org for further information.

2.3 Procedure

Parents were instructed to record a full day of their infant's natural environment using the LENA DLP on a typical day in which the mother was the primary caregiver, and another day when the father was the primary caregiver, with few noisy day outings (e.g., birthday parties, live sports matches) planned. They were shown how to use the recording device following a laboratory visit and asked to ensure that they began the at home recordings from the time that the infant awoke in the morning until the time that the infant was being placed in bed for the evening. Infants wore a LENA recorder in a specialized vest on one day with their mother, and one day with their father. As the LENA DLPs are not waterproof, the LENA recorder and vest was removed (while still recording) and placed close by, out of the infant's reach during daytime naps and baths. Parents were also given a daily logbook where they noted the infant's routine and daytime naps on the days of the LENA recordings, and could note if any outings took place like going to the grocery store or having a visitor over. All recordings were conducted on days when the families were home for more than 65% of the recording period.

2.4 Measures

For this study reports on the following measures were extracted from the LENA recording devices with LENA automatic speech recognition software:

1. Parent word count (mother primary caregiver)
2. Parent word count (father primary caregiver)
3. Quantity of conversational turn-taking episodes (mother primary caregiver)
4. Quantity of conversational turn-taking episodes (father primary caregiver)
5. Quantity of child vocalisations (mother primary caregiver)
6. Quantity of child vocalisations (father primary caregiver)

The LENA software automatically classifies conversations and noise (TV/other), obtains *word counts* (nearby male/nearby female/ infant/ distant male/distant female), and calculates the number of *conversational –turns* (child or nearby adult initiates a conversation and the other responds within 5-seconds).

3. Results

Preliminary t-tests of demographic variables indicated that there was no difference in parent age; infant birth weight, parents' education or hours spent napping during the day long recordings, all p's > .05.

T-tests were conducted to determine whether there were any differences between the quantity of adult speech on the day that the mother was the primary caregiver compared to the father, the number of conversational turns with mothers compared to fathers, and the quantity of infant vocalisations to mothers or fathers. No statistical differences were evident, all p's > .05 (see Table 2 for descriptive statistics).

Pearson correlation coefficients were produced to examine the relationship between the quantity of parental speech input and infant vocalisations, and the quantity of conversational turn-taking episodes and infant vocalisations. As shown in Table 3, the analysis revealed significant positive correlations with conversational turn-taking with the mother and number of infant vocalisations, and between conversational turn-taking with the father and number of infant vocalistions. That is, as the frequency of conversational turns increased (calculated by the LENA software as a parent's response to infant speech), the infant produced more vocalisations.

The quantity of conversational turn-taking was also positively correlated with the quantity of mothers' speech to her infant. Surprisingly, the overall quantity of speech produced by mothers and fathers was not associated with the number of words infants produced. Although it should be noted that the correlation between the amount of infant vocalisations and mothers word counts was stronger than the correlation between the amount of infant vocalisations and fathers word counts. Figure 1 illustrates the relationship between turn-taking and infant word count during interactions with mothers (top) and fathers (bottom).

Table 2. *Descriptive statistics for adult and infant word counts*

	Mean	Min	Max
Word Count mother	14674.9	9550.0	21250.0
Word Count father	14414.3	9542.0	29521.0
Turn-taking mother	428.3	277.0	784.0
Turn-taking father	379.7	194.0	509.0
Child vocalisations with mother	1631.3	820.0	2237.0
Child vocalisations with father	1598.5	678.0	2534.0

Table 3. *Pearson correlation coefficients for maternal, paternal and infant word counts, and parent-child turn-taking episodes derived from 10-14 hour long LENA recordings. ** Correlation is significant at the 0.01 level (2-tailed)*

	1	2	3	4	5	6
1. Word Count mother	1.0					
2. Word Count father	0.181					
3. Turn-taking mother	0.805**	0.064				
4. Turn-taking father	0.436	0.424	0.487			
5. Child vocalisation with mother	0.532	0.189	.816**	0.330		
6. Child vocalisation with father	0.239	0.197	0.316	.822**	0.386	1.0

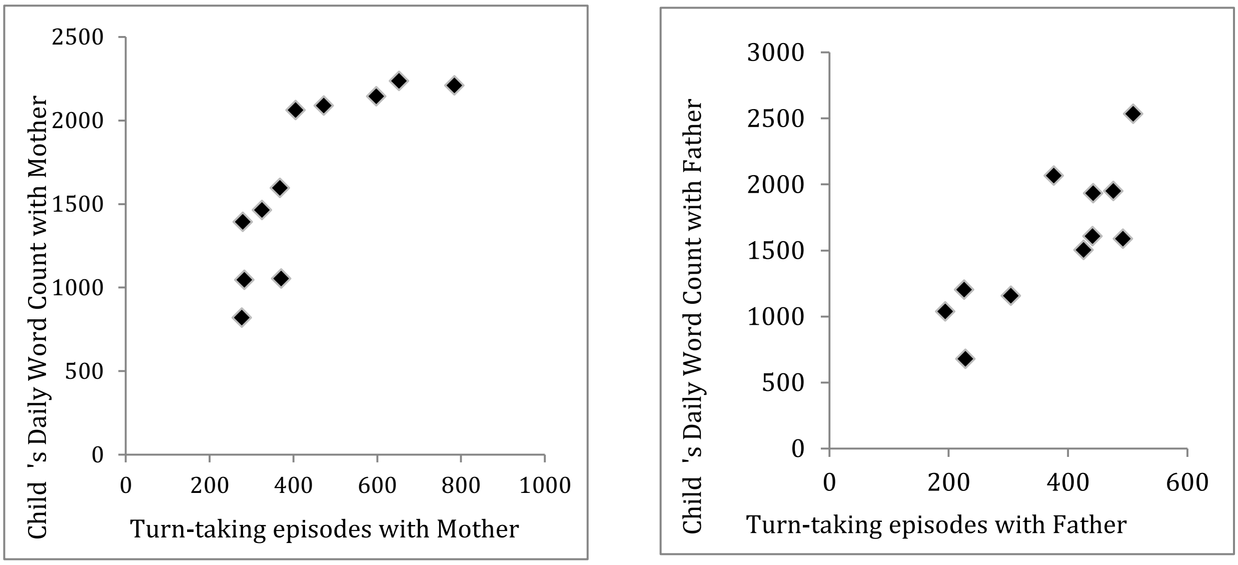


Figure 1: *Correlation between quantity of mother (top) and father (bottom) speech input and infant vocalisations during a day-long LENA recording*

4. Discussion

This paper aims to clarify the relationship between the quantity of parental speech input and early child speech output. Participants reported here consist of a subset of the participants enrolled in an ongoing longitudinal study examining qualitative and quantitative factors shapes language development who have visited the laboratory for their first appointment. The hypothesis that mothers would produce a greater quantity of speech than fathers was not supported. The quantity of speech produced by both parents was not directly related to the number of words produced by infants. Rather, quantity of conversational turn-taking between parents and child was positively related to the number of vocalisations produced by 6-month-old Australian babies. Taken together, these findings indicate that it is not simply the quantity of words spoken by mother or father is related to the number of vocalisations produced by the infant which has been shown in previous studies.

This study confirms previous findings (2-6) that infants speech output is affected by the quantity of parental speech input. Moreover, it extends our understanding by highlighting that it is not just the quantity of words that are provided, rather, it is the quantity of turns taken between parent and child that are positively related to the number of infant vocalisations at 6-months of age. This concurs with other studies that demonstrate that reciprocity between parental speech input and a child's vocalisations promotes child vocalisations and language development (5, 6). Thus it is not simply the number of words that an infant overhears on a day-to-day basis, rather it is the meaningful exchange of conversation between parent and child that promote the use of language in pre-linguistic infants.

Although we cannot determine whether the quantity of infant vocalisations encouraged more turn-taking on the part of the parent, or parent's encouraged a greater number of infant vocalisations by promoting vocal turn-taking, this study indicates that there is a critical link between parent feedback and infant vocalisations in the first 6-months of life (8). Further examination of the time series of vocalisations will help to clarify whether turn-taking episodes were instigated by the parent or infant. The results concur with evidence that the amount of quality speech input is a more potent predictor of infant language than quantity of speech input (9, 10).

This study did not find any differences in the overall quantity of words produced by mothers compared to fathers. Thus, did not support recent findings that mothers talk more than fathers with infants from birth to 7-months (7). Furthermore, infants showed no difference in the quantity of vocalisations they produced when their mother was primary caregiver, or their father was primary caregiver. However, there was a stronger positive correlation between the number of mother's words and infant vocalisations, in comparison to the corelation between father's word count and infant vocalisations. This trend will be clarified once the full sample has been tested.

This study was limited in several important ways. First, the sample was not nationally representative and therefore cannot be generalized to all Australian families. Furthermore, the data is based on a small sub-group sample of participants who had completed their initial visit in an ongoing longitudinal study examining the relationship between quantitative and qualitative measures of parental speech input and infant language development. Additional data is currently being collected to meet an adequate sample size.

Despite these limitations it is clear that it is not simply the quantity of parental speech input that predicts the extent to which infants vocalise in the first 6 months of life. Rather, it is the quantity of conversational turns between infant and parent (mothers and fathers) that promote infant vocalisations. This is an important factor that is being addressed in the longitudinal study. It is anticipated that the longitudinal follow-up of the infant's natural speech environment from the pre-linguistic to linguistic stage of development will identify how speech input may be enhanced to populations that are at-risk of language delays.

5. Conclusions

In conclusion, the present findings concur with emerging evidence (e.g., 7, 9, 10) that the quantity of reciprocal vocalizations between parent and child are a key factor in promoting language development. In contrast to other studies with pre-linguistic infants, this study did not find any difference in the amount of speech produced by mothers and fathers, or by infant's according to the sex of their interlocutor (7). These findings indicate the importance of encouraging male and female caregivers to engage in turn-taking conversations with young pre-linguistic infants. Further work is currently underway to determine whether these findings are shown in a larger sample of infants currently enrolled in a longitudinal study examining factors that facilitate early word learning.

6. Acknowledgements

This study was supported by a grant from the Marcus and Amalia Wallenburg foundation to Dr Iris-Corinna Schwarz, Dr Christa Lam-Cassettari, and Dr Christine Kitamura, and Western Sydney University honours funding to Paige Noble.

7. References

[1] Fenson, L., et., al. (1994). *Variability in early communicative development.* Monographs of the Society for Research in Child Development. 59(5), 1-173.

[2] Hart, B., & Risley, R. R. (1995). Meaningful differences in the everyday experiences of young American children. Baltimore, MD: Paul H. Brooks.

[3] Weisleder, A., & Fernald, A. (2013). Talking to children matters: Early language experience strengthens processing and builds vocabulary. *Psychological Science*, *24*(11), 2143-2152. doi:10.1177/0956797613488145

[4] Huttenlocher, J., Haight, W., Bryk, A. S., Seltzer, M., & Lyons, T. (1991). Early vocabulary growth: Relation to language input and gender. *Developmental Psychology, 27*(2), 236-248.

[5] Hsu, N., Hadley, P. A., & Rispoli, M. (2015: FirstView). Diversity matters: parent input predicts toddler verb production. *Journal of Child Language*, 1-24. doi:10.1017/S0305000915000690.

[6] Zimmerman, F. J., Gilkerson, J., Richards, J. A., Christakis, D. A., Xu, D., Gray, S., & Yapanel, U. (2009). Teaching by listening: The importance of adult-child conversations to language development. *Pediatrics*, *124*(1), 342-349. doi: 10.1542/peds.2008-2267

[7] Johnson, K., Caskey, M., Rand, K., Tucker, R., & Vohr, B. (2014). Gender differences in adult-infant communication in the first months of life. *Pediatrics, 134*(6), e1603-e1610. Doi: 10.1542/peds.2013-4289

[8] Goldstein, M. H., Schwade, J. A., & Bornstein, M. H. (2009). The value of vocalizing: five-month-old infants associate their own noncry vocalizations with responses from caregivers. *Child Development, 80*(3), 636-644.

[9] Hoff, E. (2003). The specificity of environmental influence: socioeconomic status affects early vocabulary development via maternal speech. *Child Development.* *74*, 1368–1378. [PubMed: 14552403]

[10] Rowe, M. L. (2012). A longitudinal investigation of the role of quantity and quality of child-directed speech in vocabulary development. *Child Development. 83,* 1762–1774. [PubMed: 22716950]

Noise-robust Linear Prediction Cepstral Features for Network Speech Recognition

Aadel Alatwi[1], Stephen So[1], Kuldip K. Paliwal[1]

[1]School of Engineering, Griffith University, Brisbane, QLD 4111, Australia

aadel.alatwi@griffithuni.edu.au, s.so@griffith.edu.au, k.paliwal@griffith.edu.au

Abstract

In this paper, we propose a perceptually-motivated method for modifying the speech power spectrum to obtain a set of linear prediction coding (LPC) parameters that possess good noise-robustness properties in network speech recognition. Speech recognition experiments were performed to compare the accuracy obtained from MFCC features extracted from AMR-coded speech that use these modified LPC parameters, as well as from LPCCs extracted from AMR bitstream parameters. The results show that when using the proposed LP analysis method, the recognition performance was on average 1.2% - 6.1% better than when using the conventional LP method, depending on the recognition task.

Index Terms: Linear prediction coding parameters; Network speech recognition; Automatic speech recognition

1. Introduction

Speech processing technologies are increasingly being incorporated into modern devices and applications such as Automatic Speech Recognition (ASR). This is partly because it can be simply used to provide service accessibility for clients. Many ASR applications are based on the Network Speech Recognition (NSR) approach, in what is known as a client-server model [1]. In this model, speech signals are compressed using conventional speech coders such as the GSM speech coder and transmitted to the server side, where feature extraction and the speech recognition are conducted [1]. There are two models of NSR systems: speech-based NSR (as shown in Figure 1), where the feature extraction is carried out on the reconstructed speech; and bitstream-based NSR (as shown in Figure 2), where the LPC parameters from the bitstream are converted to ASR features.

In the speech coder at the client side, the autocorrelation method [2] is typically used as the linear prediction coding (LPC) analysis technique to obtain the LP coefficients from short frames of speech, which are converted to some suitable LPC parameters, such as Log Area Ratios (LARs) or Line Spectral Frequencies (LSFs) [3]. These LP coefficients represent the power spectral envelope, which offers a concise representation of important properties of the speech signal. In noise-free environments, the performance of this LPC analysis technique is often satisfactory. However, in the presence of noise, the autocorrelation method yields a poor estimate of an all-pole model of the input speech signal [4]. This behavior results in an overall deterioration in the reconstructed speech quality, which also degrades the recognition performance at the server end [5].

This paper presents a new perceptually-inspired method of estimating LP coefficients, which we call the Smoothed and Thresholded Power Spectrum linear prediction (STPS-LP) coefficients. This method involves computing autocorrelation coefficients from a modified speech power spectrum, which are used in the autocorrelation method [2]. These LP coefficients can then be converted to LPC parameters that are compatible with current speech coders, with the added benefit of enabling noise-robust ASR features to be extracted on the server side. We have evaluated the effectiveness of the proposed method in comparison with conventional ASR features in terms of the recognition performance using both the speech-based and bitstream-based NSR approaches under clean and noisy conditions.

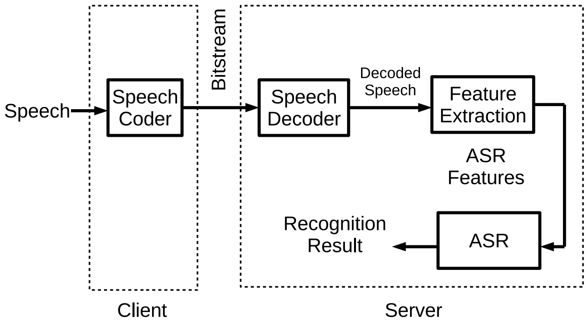


Figure 1: *Block diagram of speech-based network speech recognition (NSR).*

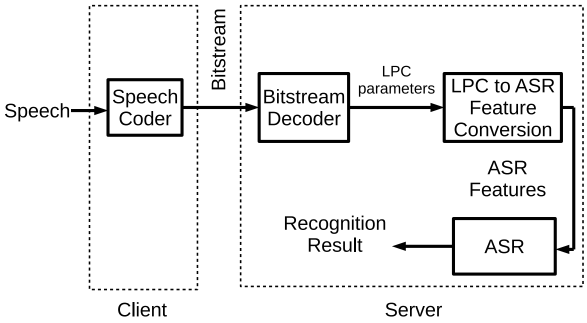


Figure 2: *Block diagram of bitstream-based network speech recognition (NSR).*

The structure of this paper is organized as follows: Section 2 explains the theory behind the proposed STPS-LP analysis method, describes the proposed algorithm, and presents the STPS-LP cepstral features at the server side. Section 3 shows the experimentally obtained results, in which we evaluate the ASR performance. Finally, we provide our conclusion in Section 4.

2. Proposed STPS-LP features for ASR

2.1. Conventional LPC analysis method

The power spectrum of a short frame $\{x(n), n = 0, 1, 2, ..., N-1\}$ of N samples of the speech signal can be modeled using an all-pole or autoregressive (AR) model [6]:

$$\hat{X}(z) = \frac{G}{1 + \sum_{k=1}^{p} a_k z^{-k}} \tag{1}$$

where p is the order of the AR model, $\{a_k,\ 1 \leqslant k \leqslant p\}$ are the AR parameters, and G is a gain factor. The parameters $\{a_k\}$ and G are estimated by solving the Yule-Walker equations [7]:

$$\sum_{k=1}^{p} a_k R(j-k) = -R(j), \quad \text{for } k = 1, 2, \ldots, p \tag{2}$$

$$G^2 = R(0) + \sum_{k=1}^{p} a_k R(k) \tag{3}$$

where $R(k)$ are the autocorrelation coefficients, which are estimated using the following formula [7]:

$$R(k) = \frac{1}{N} \sum_{n=0}^{N-1-k} x(n)x(n+k) \tag{4}$$

It can be readily shown that this AR modelling procedure of solving the Yule-Walker equations is equivalent to the autocorrelation method in linear prediction analysis [6]. In the linear prediction context, the AR parameters $\{a_k\}$ are the LP coefficients, and G^2 is the minimum squared prediction error.

The autocorrelation coefficients used in the Yule-Walker equations can also be calculated by taking the inverse discrete-time Fourier transform of the periodogram $P(\omega)$ estimate of the power spectrum [7]:

$$R(k) = \frac{1}{2\pi} \int_{-\pi}^{\pi} P(\omega) \mathrm{e}^{j\omega k} d\omega \tag{5}$$

where

$$P(\omega) = \frac{1}{N} \left| \sum_{n=0}^{N-1} x(n) \mathrm{e}^{-j\omega n} \right|^2 \tag{6}$$

This provides a way of introducing preprocessing of the periodogram to reduce the variance and improve the noise robustness prior to the computation of the LP coefficients.

2.2. Estimating perceptually motivated LPC parameters

The proposed method computes the LPC parameters in two steps: In the first step, it manipulates the periodogram estimate of the power spectrum of the speech signal with the aim of reducing the variance of the spectral estimate and removing the parts that are more affected by noise. In the second step, the autocorrelation coefficients are obtained from the processed power spectrum. The processed power spectrum is obtained through smoothing followed by thresholding operations. In the smoothing operation, as shown in Figure 3, the variance of the spectral estimate is reduced by smoothing the periodogram of the input speech signal [7] using triangular filters, which are spaced using the Bark frequency scale [8]. This non-linear smoothing operation, which is inspired by the human auditory system, results in less smoothing at low frequencies, where the high power components are located, while more smoothing is applied at the higher frequencies, where weaker spectral components are more affected by noise [9]. Following the smoothing, a thresholding operation is performed, where the influence of low signal-to-noise ratio (SNR) spectral components, which are prone to being corrupted by noise and also add unnecessary variance to the spectral estimate, are removed and replaced by the smoothed spectrum, as shown in Figure 4. As a consequence of smoothing followed by thresholding, the dominant spectral peaks are preserved because they are the least affected by noise, while the less reliable spectral valleys are discarded and replaced by a smoothed average. Hence, by improving the robustness of the power spectrum estimation, the linear prediction coefficients derived from it would have lower variance and possess better robustness in noisy environments.

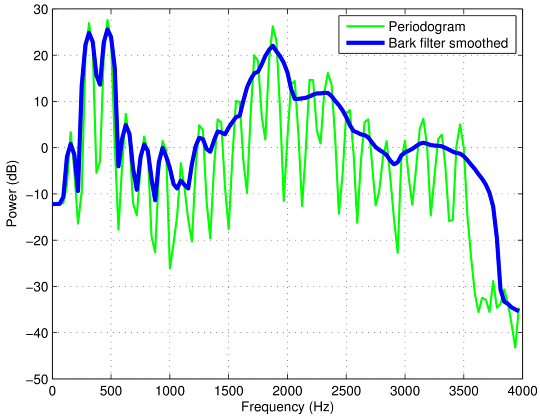


Figure 3: *Periodogram $P(k)$ and the smoothed spectrum $\bar{P}(k)$ of speech sound (vowel /e/ produced by male speaker).*

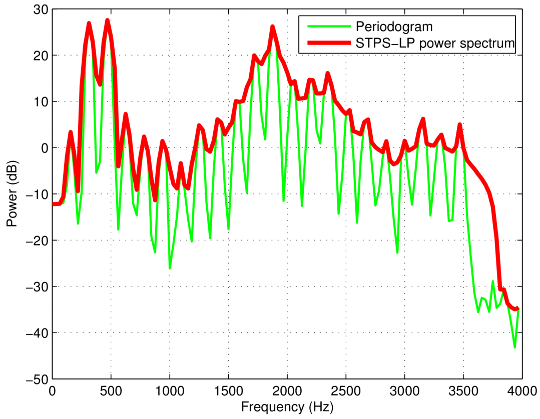


Figure 4: *Periodogram $P(k)$ and the resultant spectrum $\hat{P}(k)$ after thresholding operation of speech sound (vowel /e/ produced by male speaker).*

The proposed algorithm is described in the following steps:

Step 1: Compute the periodogram spectrum $P(k)$ of a given frame $\{x(n), n = 0, 1, 2, ..., N-1\}$ of N samples from a speech signal [7]:

$$P(k) = \frac{1}{N}\left|\sum_{n=0}^{M-1} x(n)w(n)\mathrm{e}^{-j2\pi kn/M}\right|^2, \quad 0 \leqslant k \leqslant M-1 \tag{7}$$

where $P(k)$ is the value of the estimated power spectrum at the k^{th} normalized frequency bin, M is the FFT size where $M > N$, and $w(n)$ is a Hamming window.

Step 2: Smooth the estimated power spectrum $P(k)$ using a triangular filter at every frequency sample:

$$\bar{P}(k) = \sum_{l=-L(k)}^{L(k)} K(l)P(l-k) \tag{8}$$

where $\bar{P}(k)$ is the smoothed $P(k)$, $K(l)$ is the triangular filter, and $L(k)$ is half the critical bandwidth of the triangular filter at frequency sample k. The triangular filter $K(l)$ is spaced using the Bark frequency scale, which is given by [8]:

$$Bark(f) = 13\arctan(0.00076f) + 3.5\arctan\left[\left(\frac{f}{7500}\right)^2\right] \tag{9}$$

Step 3: Using the smoothed $\bar{P}(k)$ as the threshold, $\hat{P}(k)$ is formed by retaining only spectral components that are above the threshold. This is defined as:

$$\hat{P}(k) = \begin{cases} P(k) & \text{if } P(k) \geq \bar{P}(k) \\ \bar{P}(k) & \text{otherwise} \end{cases}$$

Step 4: Compute the modified autocorrelation coefficients by taking an inverse discrete Fourier transform [7]:

$$\hat{R}(q) = \frac{1}{M}\sum_{k=0}^{M-1} \hat{P}(k)\mathrm{e}^{j2\pi kq/M} \quad , \quad 0 \leqslant q \leqslant M-1 \tag{10}$$

These autocorrelation coefficients $\hat{R}(q)$, $0 \leqslant q \leqslant p$, where p is the LPC analysis order, are then used in the Levinson-Durbin algorithm [7] to compute the linear prediction coefficients, which we call the Smoothed and Thresholded Power Spectrum linear prediction (STPS-LP) coefficients.

2.3. Cepstral features derived from STPS-LP coefficients for noise-robust speech recognition

For automatic speech recognition at the server end, the STPS-LP coefficients are extracted from the speech coding bitstream and then converted to a set of robust ASR cepstral-based feature vectors. In comparison with conventional LP cepstral coefficients (LPCCs), where the entire power spectrum is modeled by linear prediction analysis on a linear frequency scale, STPS-LP cepstral coefficients (or STPS-LPCCs) have the distinct advantage of being derived from a power spectrum that has been smoothed by an auditory filterbank and thresholded to remove low SNR spectral components. These operations reduce the influence of unreliable spectral components, which improve the feature's robustness to noise. We propose the following steps in the computation:

Step 1: Given the STPS-LP coefficients $\{a_k, k = 1,2,3,...,p\}$ and the excitation energy G^2, the power spectral estimate $P(\omega)$ is computed as follows [7]:

$$P(\omega) = \frac{G^2}{\left|1 + \sum_{k=1}^{p} a_k \mathrm{e}^{-j\omega k}\right|^2} \tag{11}$$

Step 2: Sample the power spectral estimate $P(\omega)$ at multiples of 0.5 bark scale, from 0.5 to 17.5 bark (to cover the range of 4 kHz), to give power spectral samples $\{\tilde{P}(r); r = 1,2,...,35\}$, where r is the sample number.

Step 3: Take the logarithm of each power spectral sample and compute the discrete cosine transform to produce a set of STPS-LPCCs [10]:

$$C(k) = \sqrt{\tfrac{2}{R}}\sum_{r=1}^{R} \log\tilde{P}(r)\cos\left[\tfrac{\pi}{R}\left(r - \tfrac{1}{2}\right)k\right], \quad 1 \leqslant k \leqslant N_c \tag{12}$$

where R = 35 and N_c is the desired number of cepstral coefficients.

3. Results and Discussion

In this section, a series of ASR tests were conducted to evaluate the NSR performance on speech that has been coded using LPC parameters that were derived from the conventional LP and STPS-LP coefficients in clean and noisy conditions. These ASR experiments were performed using MFCC (Mel frequency cepstral coefficients) features that were computed from the reconstructed speech (speech-based NSR); as well as using LPCC and STPS-LPCC features computed from the GSM coder parameters themselves (bitstream-based NSR). We utilized the Adaptive-Multi Rate coder (AMR) in 12.2 kbit/s mode, which is identical to the GSM Enhanced Full Rate [11]. There are three conditions that we tested:

- Baseline: training and testing on uncoded speech
- Matched: training on coded speech, testing on coded speech
- Mismatched: training on uncoded speech, testing on coded speech

In this study, all of the experiments were conducted using the DARPA Resource Management Continuous Speech Database (RM1) [12] under clean and noisy conditions. The training and test sets consisted of 3977 and 300 sentences, respectively. In all cases, the speech signal was downsampled to 8 kHz. For noisy conditions, the speech signal was corrupted by additive zero-mean Gaussian white noise at six different SNRs, ranging from 30 dB to 5 dB in 5 dB steps. The HTK toolkit [10] was used for the Hidden Markov Model (HMM) construction. The cepstral feature vector was composed of a 12 dimension base feature concatenated with delta and acceleration coefficients. Thus, the size of the feature vector was 36 coefficients. The recognition performance is represented by numerical values of word-level accuracy.

3.1. Recognition accuracy in speech-based NSR

Table 1 compares the performance of speech recognition accuracy using Mel Frequency Cepstral Coefficients (MFCCs) computed from the original speech signal without AMR coding (Column 2) and with AMR processed speech (Columns 3 - 6). The AMR speech was coded using the LPC parameters that were derived from the conventional LP and the proposed STPS-LP coefficients. Columns 3 and 4 show the results for

Table 1: *Word-level accuracies* (%) *obtained using Mel-Frequency Cepstral Coefficients derived from the original waveform and from the reconstructed speech.*

Noise Level (dB)	Baseline	Matched Models		Mismatched Models	
		LP	STPS-LP	LP	STPS-LP
Clean	95.47	94.16	94.55	93.53	93.85
30	94.46	93.57	93.65	92.12	93.18
25	92.79	92.08	92.32	91.26	91.65
20	89.58	89.07	89.27	86.17	86.99
15	77.59	72.96	75.55	68.77	71.75
10	50.31	46.11	48.53	42.40	44.79
5	16.61	14.15	15.95	12.85	13.95
Average between 5 - 30 dB	70.22	67.99	69.22	65.60	67.06

the matched condition, where the training model was computed from AMR coded speech. Columns 5 and 6 show the results for the mismatched condition, where the training model was computed from the original uncoded speech. The results show that the performance of speech-based NSR was slightly better in matched compared to mismatched models under clean condition where the speech was reconstructed using the conventional LP coefficients. This behavior did not hold in the environments of noise, especially for SNRs below 20 dB, where the performance has deteriorated in both models. On the contrary, when considering the coded speech that used the proposed STPS-LP coefficients, the provided performance is close to the baseline under almost all environmental conditions in the matched model. The recognition accuracy has improved by about 2.98% at 15 dB in comparison to AMR coding that employed the conventional method for mismatched training.

3.2. Recognition accuracy in bitstream-based NSR

Cepstral features were obtained from unquantized as well as quantized LSFs (which were derived from conventional LP and STPS-LP coefficients) that were encoded in the AMR coding bitstream. The LSF parameters (based on the conventional LP analysis method) were transformed into LP coefficients [3], and cepstral coefficients were computed using the approach described in [13] to obtain LPCCs. The proposed method that was described in Section 2.3 was used to compute STPS-LPCCs. Table 2 compares speech recognition accuracies obtained when using linear prediction cepstral features in matched conditions (training and testing based on quantized LSFs) and mismatched conditions (training based on unquantized LSFs and testing based on quantized LSFs). The results in Table 2 illustrate that, under clean conditions, there was modest improvement in the bitstream-based NSR accuracy obtained using STPS-LPCC features over LPCC features in matched and mismatched models. The STPS-LPCC features were superior to the conventional method when the speech was corrupted by white noise (that is SNR $<$ 20 dB), and in these cases the NSR performance was on an average 5.47% and 7.74% better than the conventional LPCCs in matched and mismatched models, respectively, while the baseline STPS-LPCCs was an average 6.91% better than the baseline LPCCs.

4. Conclusion

This paper has presented a new method of estimating LP coefficients that are designed to exploit the non-linear spectral selectivity of the human auditory system. These LP coefficients and their associated LPC parameters are fully compatible with industry-standard LP-based speech coders. Through smoothing and thresholding, low energy spectral components that are more vulnerable to being corrupted by noise are discarded, resulting in lower estimation variance and subsequently improved noise robustness in ASR. The speech recognition performance of the STPS-LP coefficients, in comparison with conventional LP coefficients, was investigated for two NSR scenarios. The recognition accuracy improved slightly when using MFCC features derived from speech that was coded using STPS-LP-based parameters for all SNRs and in both matched and mismatched conditions. In the bitstream NSR scenario, STPS-LPCC features computed from the bitstream parameters resulted in higher recognition accuracies, especially at lower SNRs. The results demonstrated the improved noise-robustness of the STPS-LP coefficients.

Table 2: *Word-level accuracies* (%) *obtained using Cepstral Coefficients derived from the LSF parameters that were transformed into the corresponding LPC coefficients.*

Noise Level (dB)	Baseline		Matched Models		Mismatched Models	
	LPCCs	STPS-LPCCs	LPCCs	STPS-LPCCs	LPCCs	STPS-LPCCs
Clean	91.98	93.08	90.34	91.75	88.74	90.77
30	88.62	90.07	87.49	89.90	84.51	87.33
25	86.16	88.35	85.02	87.25	79.66	85.33
20	83.26	86.66	81.11	84.08	76.07	80.95
15	75.28	81.07	73.72	78.14	64.84	73.29
10	57.88	66.60	56.14	62.42	47.91	56.55
5	30.07	36.30	29.48	35.18	24.64	30.78

5. References

[1] S. So and K. K. Paliwal, "Scalable distributed speech recognition using gaussian mixture model-based block quantisation," *Speech communication*, vol. 48, no. 6, pp. 746–758, 2006.

[2] J. Makhoul, "Linear prediction: A tutorial review," *Proceedings of the IEEE*, vol. 63, no. 4, pp. 561–580, 1975.

[3] W. B. Kleijn and K. K. Paliwal, *Speech coding and synthesis*. New York, NY, USA: Elsevier Science Inc., 1995.

[4] S. M. Kay, "The effects of noise on the autoregressive spectral estimator," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 27, no. 5, pp. 478–485, 1979.

[5] A. Trabelsi, F. Boyer, Y. Savaria, and M. Boukadoum, "Improving lpc analysis of speech in additive noise," in *Circuits and Systems, 2007. NEWCAS 2007. IEEE Northeast Workshop on*. IEEE, 2007, pp. 93–96.

[6] J. Makhoul, "Spectral linear prediction: properties and applications," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 23, no. 3, pp. 283–296, 1975.

[7] M. H. Hayes, *Statistical digital signal processing and modeling*. John Wiley & Sons, 2009.

[8] H. Fletcher, "Auditory patterns," *Reviews of modern physics*, vol. 12, no. 1, p. 47, 1940.

[9] B. Moore, *An Introduction to the Psychology of Hearing*. Academic Press, 1997.

[10] S. J. Young, G. Evermann, M. J. F. Gales, T. Hain, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. C. Woodland, *The HTK Book, version 3.4*. Cambridge, UK: Cambridge University Engineering Department, 2006.

[11] ETSI, *ETSI TS 126 090 Digital cellular telecommunications system (Phase 2+); Universal Mobile Telecommunications System (UMTS); AMR speech Codec;Transcoding Functions (3GPP TS 26.090 version 7.0.0 Release 7)*. Tech. Rep., 2007.

[12] W. M. Fisher, G. R. Doddington, and K. M. Goudie-Marshall, "The darpa speech recognition research database: specifications and status," in *Proc. DARPA Workshop on speech recognition*, Feb 1986, pp. 93–99.

[13] L. Rabiner and B.-H. Juang, *Fundamentals of speech recognition*. Prentice hall, 1993.

Formant dynamics and durations of *um* improve the performance of automatic speaker recognition systems

Vincent Hughes, Paul Foulkes, Sophie Wood

Department of Language and Linguistic Science, University of York, UK

vincent.hughes@york.ac.uk, paul.foulkes@york.ac.uk

Abstract

We assess the potential improvement in the performance of MFCC-based automatic speaker recognition (ASR) systems with the inclusion of linguistic-phonetic information. Likelihood ratios were computed using MFCCs and the formant trajectories and durations of the hesitation marker *um,* extracted from recordings of male standard southern British English speakers. Testing was run over 20 replications using randomised sets of speakers. System validity (EER and C_{llr}) was found to improve with the inclusion of *um* relative to the baseline ASR across all 20 replications. These results offer support for the growing integration of automatic and linguistic-phonetic methods in forensic voice comparison.

Index Terms: forensic voice comparison, automatic speaker recognition, hesitation markers, formant dynamics

1. Introduction

Forensic voice comparison (FVC) accounts for the majority of casework conducted by forensic speech scientists. FVC typically involves the comparative analysis of speech samples of a known suspect (e.g. police interview) and an unknown offender (e.g. covert drug deal). In such cases, it is the role of the expert to evaluate the strength of the speech evidence under the competing propositions of the prosecution (i.e. the suspect and the offender are the same person) and the defence (i.e. the suspect and the offender are different people).

Two sets of methods are commonly used in FVC: auditory-acoustic (linguistic-phonetic) analysis and automatic speaker recognition (ASR). These methods have largely developed independently. However, a growing body of research focuses on the integration of the methods to improve the performance of FVC systems. [1] and [2] investigated the performance of a generic Mel frequency cepstral coefficient (MFCC)-based ASR system when fused with formant and tone (f0) trajectories of vowels in standard Chinese. The results show that the fusion of linguistic-phonetic and ASR systems improves performance above the baseline ASR. However, smaller improvements in validity were obtained with mobile phone recordings. The authors therefore conclude that labour-intensive linguistic-phonetic analysis may be unwarrented in FVC casework. [3] present promising results resolving the false acceptances produced by an i-vector-based ASR using voice quality analysis. The move towards an integrated approach is also highlighted by the inclusion of a human-assisted ASR (HASR) element within the NIST evaluations in 2010 [4]. Further, the use and acceptance of ASRs in conjunction with linguistic-phonetic analysis in casework is increasing, with labs in Germany and Sweden providing conclusions based on combinations of analyses.

In [5] we presented the results of likelihood ratio (LR)-based testing using combinations of different spectral and temporal features extracted from the hesitation markers *uh* and *um*. Hesitation markers are thought to be good speaker discriminants since they occur frequently, are less susceptible to coarticulation than lexical vowels, and display less within-speaker variability since speakers have little conscious control over their production [6,7]. In [5], testing was conducted using single recordings from a set of 60 young male speakers of standard southern British English (SSBE) [8]. Different combinations of input variables for each hesitation type were analysed and compared in terms of strength of evidence and system performance. The best performing system used the F1, F2, and F3 trajectories of the vocalic portion of *um* fitted with quadratic polynomials, together with vowel and nasal durations. This system achieved an equal error rate (EER) of 4.08% and a Log LR cost (C_{llr}) [9] of 0.12. A number of general findings also emerged from these tests. First, *um* consistently performed better than *uh*. Second, the inclusion of information from the first three formants outperformed any individual formant or combination of two formants. Third, modelling the formant trajectories of *um* dynamically (i.e. with multiple measurements across the duration of the vowel) outperformed static midpoint analyses. However, for *uh*, midpoint input outperformed dynamics. Finally, the inclusion of durations consistently improved system performance.

The present study expands on the promising results of [5] to assess the potential additional value of combining MFCC-based ASR systems with the best performing hesitation system, i.e. the formant dynamics and durations of *um*. As in [1] and [2], the ASR acts as a baseline system against which the individual and fused systems are compared. Performance is evaluated in terms of both EER and C_{llr}. This study builds on [5] in a number of ways. The same corpus is used, but two recordings of each speaker in separate forensically relevant tasks are analysed. This provides a more realistic estimation of the within-speaker variability in FVC casework, and therefore a more realistic representation of the performance of the systems under casework conditions. The analysis includes more data per speaker than in [5]. Finally, multiple replications of the same experiment are conducted using randomised sets of speaker.

2. Methodology

2.1. Recordings

Data were drawn from the Dynamic Variability in Speech (DyViS) corpus [8]. DyViS contains male speakers of SSBE aged 18-25. Recordings of Tasks 1 and 2 were used. Task 1 involves a mock police interview in which the participant is questioned about a crime. Task 2 involves an information exchange task conducted over the telephone between the participant and an accomplice. For this study the high quality, near-end studio recordings of both tasks were used. Both tasks are around 15 minutes in duration. In their design, DyViS tasks 1 and 2 capture the situational differences across recordings (e.g. interlocutor, topic, Lombard speech due to telephone transmission) typical in real FVC casework. The tasks were recorded in separate sessions on the same day. There was thus some time between the two sessions.

2.2. Feature extraction

2.2.1. Linguistic-phonetic system

The linguistic-phonetic system consisted of quadratic polynomial coefficients derived from the F1 to F3 trajectories of the vocalic portion of *um*, as well as vowel and nasal durations. PRAAT TextGrids containing manually segmented tokens of *um* were already available for 88 of the 100 speakers for Task 1. *um* tokens from the Task 2 recordings were also segmented for the same 88 speakers. For both tasks, F1 to F3 values were extracted at +10% steps across each vowel, tracking between five and six formants within a range of 0 to 5kHz. Vowel and nasal durations were also extracted.

The raw data were inspected visually and obvious measurement errors removed. Missing values were replaced with the mean of the values for the adjacent steps. A series of heuristics were then applied to remove less obvious errors. Data points outside specific ranges were removed: 250-900Hz for F1, 900-1900Hz for F2, and 1900-3200Hz for F3. Univariate outliers were calculated based on the group mean at each +10% step. Values of greater than ±3.29 standard deviations from the mean were removed. Where possible, missing values were again replaced with the mean of adjacent values. Finally, formant trajectories were fitted with quadratic polynomials, generating three coefficients per formant.

Speakers with fewer than 20 tokens per sample were removed, leaving a data set of 63 speakers with between 20 and 49 tokens per sample (mean=38). Although the number of tokens per speaker may appear unrealistically large relative to real case data, the availability of large amounts of data is increasingly common in FVC casework, especially in high profile cases conducted over many months or years.

2.2.2. Automatic system

A generic MFCC-based Gaussian Mixture Model-Universal Background Model (GMM-UBM) system [10] was used as a baseline against which to assess the performance of the *um* and fused systems. Pre-processing was conducted to isolate the speech-active portion of each sample. Recordings were edited manually to remove overlapping speech, interlocutor speech, clicks and background noise. Automatic clipping detection was then run, and clipped sections removed. Finally, voice activity detection was performed using the `voicebox` toolkit in MATLAB to remove silences greater than 100ms. Utterances were then concatenated into a single sample.

The audio were resampled at 10kHz (frequency range = 0-5000Hz) and MFCCs were extracted using the `rastamat` toolkit in MATLAB. A pre-emphasis filter (coefficient value = 0.97) was applied to each sample. Samples were then divided into a series of frames using a 20ms hamming window shifted at 10ms across the duration of the sample, i.e. with 50% overlap between adjacent frames. A Mel filterbank consisting of triangular filters was applied to the power spectrum of the signal for each frame. The energy in each filter was summed and logged, and the log filterbank fitted with a discrete cosine transform (DCT). The coefficients from the DCT are MFCCs. From each frame, 16 MFCCs were extracted. 16 delta and 16 delta-delta coefficients were also appended to the feature vector for each frame. Following [11], data from three frames before and after utterance boundaries were removed.

2.3. Likelihood ratio (LR)-based system testing

Likelihood ratios (LRs) was used to evaluate the performance of the individual and fused systems. The LR is expressed as:

$$LR=\frac{p(E\,|\,H_p)}{p(E\,|\,H_d)}, \tag{1}$$

where p is probability, E is evidence, H_p is the prosecution proposition and H_d is the defence proposition. The numerator of the LR is equivalent to the similarity between the suspect and offender samples, while the denominator is equivalent to the typicality (or distinctiveness) of the offender sample relative to patterns in the relevant population [12].

2.3.1. Feature-to-score stage

From the 63 available speakers, 60 were identified and randomly divided into sets of 20 training speakers, 20 test speakers, and 20 reference speakers. Same- (SS; 20) and different-speaker (DS; 190) comparisons were conducted using the training and test sets separately, with the reference set used to calculate typicality. Each comparison generates a LR-like score. Scores for the *um* system were computed using a MATLAB implementation [13] of Aitken and Lucy's multivariate kernel density (MVKD) formula [14]. MVKD models the suspect data with a normal distribution and the reference data with kernel density made up of equally weighted Gaussians for each reference speaker

GMM-UBM scores for the MFCC system were computed using the MSR Identity Toolbox [15]. A 512 Gaussian UBM was trained on data from the 20 reference speakers. Suspect samples for each development and test speaker were created using maximum a posteriori (MAP) adaptation. The suspect data were first modelled as a 512 Gaussian GMM. The GMM is parameterised using the means, variances and weights of the Gaussians. For each suspect, a copy of the UBM is made and then adapted towards the means, variances and weights from the suspect data. This is then used as the suspect model. The score (s) for each suspect-offender comparison is then:

$$s=\frac{1}{T}\sum_{i=1}^{T}\log(p(x_i\,|\,\lambda_{sus})-p(x_i\,|\,\lambda_{bkg})) \tag{2}$$

where T is the number of observations in the offender data, x_i is the offender value, λ_{sus} is the suspect model and λ_{bkg} is the background (reference) model.

2.3.2. Score-to-LR stage

The *um* and MFCC systems were initially analysed separately. For each system, calibration coefficients were calculated from the scores for the training data using logistic regression. The calibration coefficients were then applied to the scores for the test data to produce sets of calibrated log LRs (LLRs). The systems were also combined using logistic-regression fusion. In all cases, calibration and fusion coefficients were calculated using a robust MATLAB implementation [16] of scripts from Brümmer's FOCAL toolkit [17].

2.3.3. *System evaluation and replication*

The validity of the systems was evaluated using EER and C_{llr} [9]. EER represents the threshold-independent point at which the percentage of false hits (DS providing SS evidence) and misses (SS providing DS evidence) is equal. In this way EER is based on categorical, accept-reject decisions. C_{llr} is a cost function which penalises the system for the magnitude of contrary-to-fact LLRs, such that high magnitude contrary-to-fact LLRs are penalised more heavily than contrary-to-fact LLRs around threshold. The closer the C_{llr} to zero, the better the validity of the system. Testing was repeated using quasi Monte Carlo simulations. 20 different randomised sets of training, test, and reference data were created, and patterns compared across replications.

3. Results

3.1.1. *Individual systems*

Table 1 displays the mean and range of validity values for the MFCC and *um* systems across the 20 replications. In 17 of the 20 replications the ASR system outperformed the linguistic-phonetic system. The ASR systems produced a mean EER of 2.57% compared with 4.83% for the *um* systems, and a mean C_{llr} of 0.144 compared with 0.261 for the *um* systems.

Table 1. *Mean and range (max-min) of C_{llr} and EER (%) values for the um and MFCC systems across 20 replications.*

System	**C_{llr}** *Mean*	**C_{llr}** *Range*	**EER** *Mean*	**EER** *Range*
um	0.261	0.751	4.83	8.68
MFCC	0.144	0.526	2.57	5.13

Nonetheless, the results for *um* are extremely promising. First, *um* outperformed the ASR system in three replications, despite the ASR system using information from the entire speech-active portion of the sample. Second, two of the *um* systems outperformed the system in [5] (where EER=4.08% and C_{llr}=0.12) despite the use of separate suspect and offender samples. The remaining 18 replications produced validity very close to that produced in [5]. This suggests that *um* is relatively robust against the type of stylistic variation commonly found across in FVC casework.

For both the linguistic-phonetic and ASR systems, the variability in validity as a function of the configuration of speakers in the training, test, and reference sets is relatively large. At least for C_{llr}, this is, in part, due to two replications which provided atypically poor validity relative to the other replications. However, even excluding these replications the range of validity values is large. The implications of this are discussed in 4.

3.1.2. *Fused systems*

Figures 1 (C_{llr}) and 2 (EER) display the validity of each of the baseline ASR and fused systems, indicating the direction and magnitude of the change in performance with the addition of the *um* system. The C_{llr} of the fused systems was found to be consistently lower across the 20 replications. The absolute improvement in C_{llr} ranged from 0.003 to 0.43, with mean improvement of 0.09. In terms of percentage improvement, the addition of *um* reduced C_{llr} by between 8.7% and 89.9% relative to the baseline ASR systems. The largest improvement in performance was found for the ASR systems with inherently high C_{llr}. For replication 15, the fusion with the *um* system reduced C_{llr} from 0.55 to 0.12. For the ASR systems with inherently better C_{llr} (i.e. closer to 0), the magnitude of the improvement in the fused system was predictably smaller.

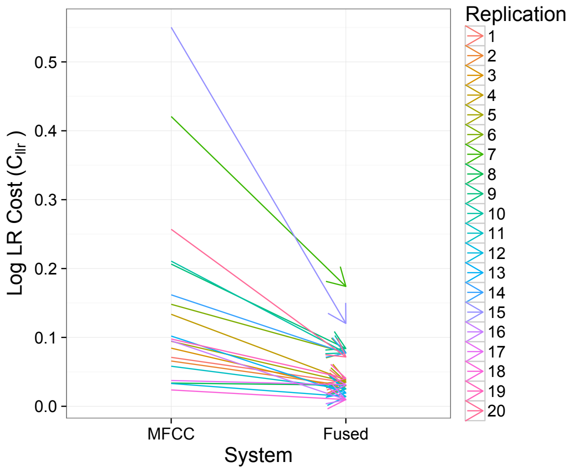


Figure 1: *C_{llr} values for the MFCC-only and fused systems across all 20 replication.*

Similar patterns were found for EER. With the exception of the three ASR systems which produced 0% EER, the remaining 17 fused systems produced lower EER than the baseline system. The absolute improvement in EER ranged from 0.26% to 5.13% (in this replication bringing EER for the fused system down to 0%), with mean improvement of 2.58%.

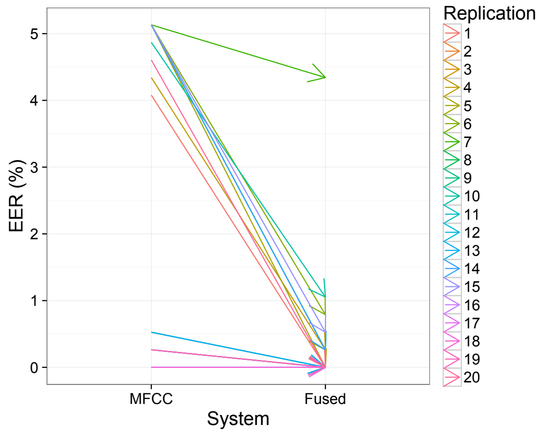


Figure 2: *EER (%) values for the MFCC-only and fused systems across all 20 replication.*

4. Discussion

4.1.1. *ASR vs. linguistic-phonetic systems*

The individual ASR and linguistic-phonetic systems performed extremely well across the tests conducted in this study, producing mean EER values of less than 5% and mean C_{llr} values of less than 0.27. In 17 of the 20 replications, the ASR system outperformed the linguistic-phonetic system, with the ASR systems optimally achieving an EER of 0% and a C_{llr} of 0.02. The performance of the *um* systems was also very good, optimally achieving an EER of 0.26% and a C_{llr} of 0.08. The extent to which the ASR systems outperformed the

Table of Contents for this manuscript

linguistic-phonetic systems is not as great as may be expected, given the considerably larger portion of the recording analysed using the ASR. The results for *um* in 3.1.1. compare very well with previous studies which have considered the performance of formant trajectory-based linguistic-phonetic systems [18,19]. Together with [5], the results offer further support for the value of filled pauses as features in FVC cases. However, somewhat poorer validity for all systems is expected when using more forensically realistic materials, incorporating a greater degree of non-contemporaneity and channel mismatch.

As shown in Table 1, however, for both forms of input the range of variability in system validity across replications is relatively large. This is purely random variation as a function of the particular speakers that make up the training, test, and reference data sets. The speakers themselves all performed the same tasks in the same way and are demographically well matched. The variability across replications is an important issue for FVC evidence, as it may have a considerable effect on the validity of the system presented to the court and the resulting strength of evidence. In the interests of transparency and objectivity, it may be necessary to perform similar replications to assess the sensitivity of system output in real FVC casework. It may then be possible for the expert to present a range of potential system validity values to the court (in the form of a credible interval).

4.1.2. Individual vs. fused systems

Despite *um* producing poorer system validity than the baseline ASRs, very promising improvement was found when the two systems were fused. Improvements in C_{llr} were found across all 20 replications. The mean absolute improvement in C_{llr} was 0.09, equivalent to a mean decrease in C_{llr} of 58.1% relative to the baseline system. Maximally, the fusion of the two systems reduced C_{llr} by 89.9%. Such improvement is considerably greater than that reported in [1]. Improvements in EER were found for 17 of the 20 replications. The three exceptions were the baseline ASR systems which were already performing at ceiling for EER (i.e. they produced 0% EER individually and 0% EER when fused with *um*).

This suggests that the speaker-discriminatory information encoded within the formant dynamics and durations of *um* may be orthogonal to that encoded within the MFCCs and derivatives. Further, the combined systems benefit from the fact that, as well as the input data being potentially independent, the speaker-discriminatory power of both systems independently was very good. This leads to almost complete separation of SS and DS pairs when fused. These results highlight the potential value of informed linguistic-phonetic analysis in FVC, and the importance of considering multiple variables (of different types) in any analysis. However, based on [2], the magnitude of the improvement in such fused systems over baseline ASRs may be less when using more forensically realistic data.

5. Conclusions

This study has shown that the performance of an MFCC-based FVC system can be improved, in some cases considerably, by incorporating the formant trajectories and durations of the vocalic portion of the hesitation marker *um*. These results highlight the value of informed linguistic-phonetic analysis in FVC, and support the move towards integrating the best elements of different methods in order to improve the validity and reliability of FVC evidence presented to courts. Future work will consider the additional benefit of linguistic-phonetic analysis to more state-of-the-art, i-vector ASR systems.

6. Acknowledgements

This research was partly funded by an ESRC PhD Scholarship to Vincent Hughes, a University of York Annual Fund Masters bursary to Sophie Wood, and by the AHRC grant *Voice and Identity* (AH/ M003396/1).

7. References

[1] Zhang, C. and Enzinger, E., “Fusion of multiple formant-trajectory- and fundamental-frequency-based forensic-voice-comparison systems: Chinese /ei1/, /ai2/, and /iau1/”, Proc. ICA - POMA 19, 2013.

[2] Zhang, C. et al., “Effects of telephone transmission on the performance of formant trajectory-based forensic voice comparison - female voices”, Speech Communication 55(6), 796-813, 2013.

[3] Gonzalez-Rodriguez, J. et al., “What are we missing with i-vectors? A perceptual analysis of i-vector-based falsely accepted trial, Proc. Odyssey, pp. 34-40, 2014.

[4] Greenberg, G. et al., “Human assisted speaker recognition (HASR) in NIST SRE2010”, Proc. Odyssey, 180-185, 2010.

[5] Anonymous, “Strength of forensic voice comparison evidence from the acoustics of filled pauses”, IJSLL 23(1):99-132, 2016.

[6] Tschapse, N., Trouvain, J., Bauer, D. and Jessen, M., “Idiosyncratic patterns of filled pauses”, IAFPA Conference, Marrakesh, Morocco, 2005.

[7] Jessen, M., “Forensic phonetics”, *Lang. Ling. Compass* 2(4):671-711.

[8] Nolan, F., McDougall, K., de Jong, G. and Hudson, T., “The DyViS database: style-controlled recordings of 100 homogeneous speakers for forensic phonetic research”, IJSLL 16(1):31-57, 2009.

[9] Brümmer, N. and du Preez, J., “Application-independent evaluation of speaker detection”, Computer Speech and Language 20(2-3):230-275, 2006.

[10] Reynolds, D. et al., “Speaker verification using adapted Gaussian mixture models”, Digital Signal Processing 10:19-41, 2000.

[11] 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”, Sci. Jus., 56(1):42-57, 2016.

[12] Aitken, C.G.G. and F. Taroni, Statistics and the Evaluation of Evidence for Forensic Scientists (2^{nd} ed), Wiley, 2004.

[13] Morrison, G.S., “MATLAB implementation of Aitken and Lucy’s (2004) forensic likelihood-ratio software using multivariate-kernel-density estimation, 2007, Online: http://geoff-morrison.net/#MVKD.

[14] Aitken, C.G.G. and Lucy, D., “Evaluation of trace evidence in the form of multivariate data”, Applied Statistics 54:109-122, 2004.

[15] Sadjadi, S.O., Slaney, M. and Heck, L., “MSR Identity toolbox v1.0: a MATLAB toolbox for speaker-recognition research”, IEEE, 2013.

[16] Morrison, G.S., “Robust version of train_llr_fusion.m from Niko Brümmer’s FOCAL Toolkit”, 2009, Online: http://geoff-morrison.net/#TrainFus.

[17] Brümmer, N., “ The FOCAL Toolkit”, Online: http://niko.brummer.googlepages.com/

[18] Rose, P., “Bernard’s 18 – vowel inventory size and strength of forensic voice comparison evidence”, Proc. SST, pp. 30-33, 2010.

[19] Rose, P., “Forensic voice comparison with monophthongal formant trajectories – a likelihood ratio-based discrimination of ‘schwa’ vowel acoustics in a close social group of young Australian females”, Proc. ICASSP, pp. 4819-4823, 2015.

Eigenfeatures: An alternative to Shifted Delta Coefficients for Language Identification

Sarith Fernando [1, 2], *Vidhyasaharan Sethu*[1], *Eliathamby Ambikairajah*[1, 2]

[1]School of Electrical Engineering and Telecommunications, UNSW Australia
[2]ATP Research Laboratory, National ICT Australia (NICTA), Australia
sarith.fernando@unsw.edu.au

Abstract

Almost all of the current LID systems use Shifted Delta Coefficients (SDC) as the temporal information features, in addition to spectral based features. However, using normalisation techniques on SDC features to make them more robust to noise or channel effects also tends to distort some of the language specific information. In this paper, we propose Eigenfeatures (EF) as an alternative to SDCs to capture temporal information while being more robust to any distortion caused by normalisation. Experimental results based on NIST LRE 2015 database shows that the Eigenfeatures are a better alternative to SDCs in the context of language recognition.

Index Terms: Shifted Delta Coefficients, Eigenfeatures, Post Log-Likelihood Ratios, Deep Bottleneck Features

1. Introduction

Feature extraction is one of the most important steps in any Language identification system [1]. Typically frame based features such as Mel Frequency Cepstral Coefficients (MFCC) [2] and Perceptual linear predictive coefficients (PLP) [3] are used and more recently, Phone Log likelihood Ratio (PLLR) [4] and deep neural network based Bottleneck Features (BNF) [5] have also been identified as effective features for language identification.

In order to capture longer term temporal context, Shifted Delta Coefficients (SDC) [6] concatenated with acoustic and phonotactic features [7, 8] are commonly employed. These coefficients have been shown to be better than just using the acoustic and phonotactic features on their own or in combination with delta and delta-delta features for language identification systems [1]. Recently, a new 'local variability feature' [9] has gained interest in speaker recognition tasks and also in image processing applications [10] and have been used to capture short term temporal information or regional variations. These local variability features have been shown to be comparable to delta and delta-delta coefficients in speaker verification tasks [9].

In this paper, we propose the use of Eigenfeatures to capture short term temporal information in place of SDCs in conjunction with the recently introduced PLLR and BNF features. Experiments carried out on NIST 2007 and NIST 2015 datasets demonstrate that the use of Eigenfeatures is a better alternative to shifted delta coefficients for language identification system.

Generally, in language identification, there is a need for features that represent language specific information which are robust to noise and channel effects. The most common way of achieving this is through suitable feature normalisation techniques [11, 12]. However, the use of normalisation techniques generally tends to make the covariances of all features with utterances of all languages to be more similar to each other [9]. This, in turn, may lead to a loss of some language specific information. Using experimental results, we show that the Eigenfeatures, which are calculated from short term covariance matrices, are less susceptible to normalisation.

2. Proposed Eigenfeature extraction

Eigenfeatures are estimated from frame based features within overlapping windows of multiple frames as shown in Figure 1. The Eigenfeatures are characterised by three main parameters, N-P-K. Where, N represents the dimensionality of the underlying frame based feature vectors, P is the number of frames contained in the window over which the Eigenfeatures are estimated and K is the number of Eigenvectors that are concatenated to form the Eigenfeature vector.

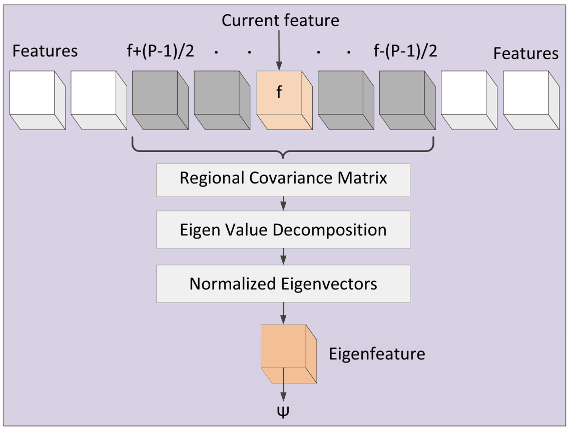


Figure 1: *Plot of Eigenfeature extraction processing starting from root features.*

Let $[X]_{P\times N}$ denote, $N-$ dimensional feature vectors corresponding to P frames in a window R. The regional covariance matrix, C_R, of the features corresponding to this window R can then be computed as,

$$C_R = \frac{1}{P-1}\sum_{t=1}^{P}(x_t - \mu)(x_t - \mu)^T \qquad (1)$$

where, x_t denotes the feature vector corresponding to the t^{th} frame and μ is the mean over the P frames.

Then Eigenvalue decomposition was then performed on this regional covariance C_R as,

$$C_R = VSV^{-1} \qquad (2)$$

where V is the $N \times N$ Eigenvector matrix and S is the $N \times N$ diagonal matrix which consists of the corresponding

Eigenvalues. Finally, each Eigenvector $\boldsymbol{v'}_i$ is normalised as follows,

$$\boldsymbol{v}_i = \frac{\boldsymbol{v'}_i \cdot s_i}{\sum_{j=1}^{N} s_j} \qquad (3)$$

where $\boldsymbol{v_i}$ is the normalised Eigenvector. Finally, the NK – dimensional Eigenfeature vector, $\boldsymbol{\Psi}$, is formed by concatenating the eigenvectors correspond to K largest eigenvalues. In our experiments we have used N-5-1 configuration where N was 59 and 42 for Phone Log Likelihood Ratios (PLLR) and Bottleneck Features (BNF) respectively.

3. System Description

A block diagram of the LID system used in our experiments reported in this paper is shown in Figure 2. This system is used to compare the Eigenfeatures (section 2) as an alternative to shifted delta coefficients [1] . The rest of the system is a typical language identification system comprising of an i-vector-GPLDA back-end with either Phone Log-Likelihood Ratio or Bottleneck features employed in the front-end. In addition, length normalisation and LDA were applied on the i-vectors prior to Gaussian Probabilistic Linear Discriminant Analysis (GPLDA) scoring [13]. It should be noted that the i-vectors corresponding to each front-end (PLLR or BNF) are estimated using a UBM and a T-matrix specific to that front-end.

3.1. PLLR feature Extraction

The Phone Log-Likelihood Ratio (PLLR) feature vectors [4] from each frame were computed using phone posteriors estimated from speech utterances using a Hungarian phone decoder developed by the Brno University of Technology [14]. Following the estimation of phone posteriors, a Voice Activity Detector (VAD) [4] was used to remove non-speech frames from each utterance. Finally, the state posteriors corresponding to each phoneme were summed together and the 3 non-phonetic units were combined into one single unit which led to 59 dimensional PLLR values. PLLR features were computed as,

$$LLR_i(t) = log\frac{p_i(t)}{\frac{1}{N-1}(1-p_i(t))} \quad ,i = 1, \dots, N \qquad (4)$$

where, $LLR_i(t)$ represents the log-likelihood ratio of the i^{th} phoneme and $p_i(t)$ represents the posterior probability of the i^{th} phoneme corresponding to frame t. The N dimensional vectors (59 dimension in our case) corresponding to each frame were then referred to as PLLR feature vectors (or PLLRs).

3.2. BNF feature extraction

The overall Deep Neural Network (DNN) architecture of Bottleneck feature extraction process is shown in Figure 3. BNF were extracted using a Deep Neural Network [5] trained on MFCC features. The Network has trained on 300 hours of Switchboard 1 data as defined in Kaldi example “tri4a” [15]. DNN consists of 5 layers each with 1024 nodes except at the bottleneck layer at layer four. All of these layers used ‘tanh’ activation function with the exception of the bottleneck layer. The bottleneck layer comprised of 42 nodes using a linear activation function. After extracting Bottleneck Features, Vector Quantization Voice Activity Detection (VQ-VAD) was used.

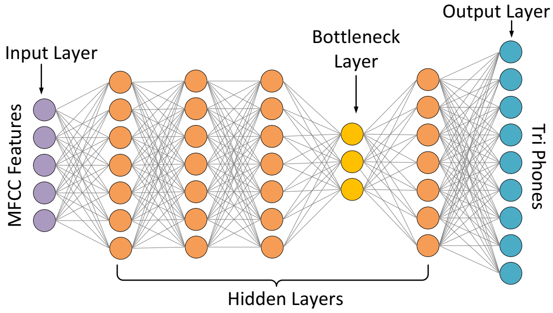


Figure 3: *Block diagram of DNN architecture for BNF extraction process*

3.3. I-Vector Gaussian PLDA

The Total Variability matrix was trained as in [16]. This is based on the language-independent Universal Background model (UBM) consisting of 1024 Gaussian components trained on 118 dimensional PLLR coefficients or on 84 dimensional BNF concatenated with either SDCs or EF separately. Only half of the data available from target languages was used for UBM training and Total Variability matrix was trained on all language data. The i-vector dimension was selected as 400 according to the previous experiments conduct on NIST 2007 dataset [13] and computed

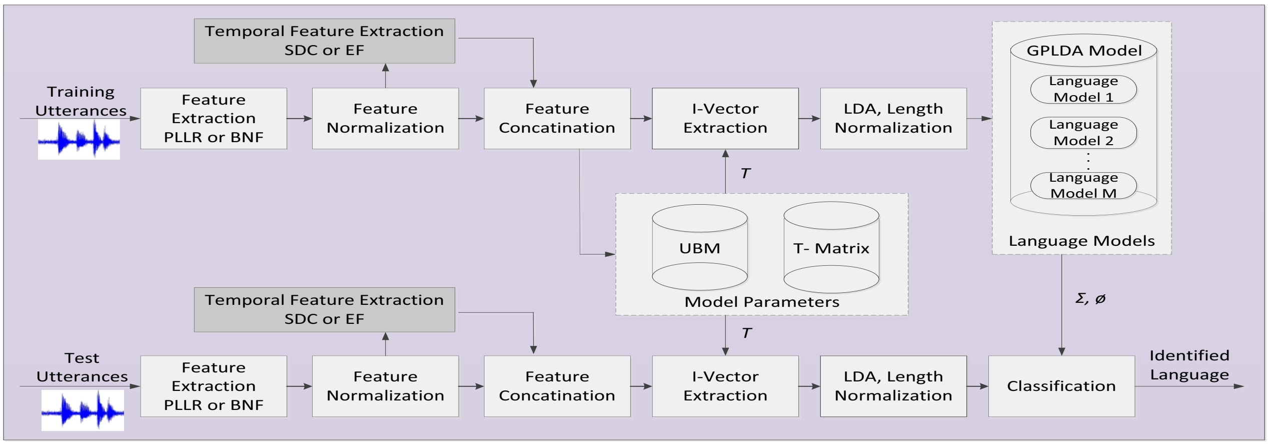


Figure 2: *Block diagram of experimental setup*

as,

$$\boldsymbol{M} = \boldsymbol{m} + T\boldsymbol{w} \qquad (5)$$

where $\boldsymbol{M}$ is the utterance dependent GMM mean supervector, $\boldsymbol{m}$ is the language independent mean super vector, T is the total variability matrix and $\boldsymbol{w}$ is the low dimension i-vector.

In this study, a Gaussian Probabilistic Linear Discriminant Analysis (GPLDA) was used as a classifier to make the final decision based on the i-vectors. This can be denoted by "Language models" block in Figure 2. The GPLDA back-end has shown to be effective, in LID tasks [13]. Typically, GPLDA based systems use length normalization of i-vectors to overcome their non-Gaussianity [14]. In the GPLDA approach, the i-vectors are represented by a generative model given as log likelihood ratio between the same $H1$ versus different $H0$ language model hypothesis,

$$s(u,v) = log\frac{p(u,v|H1)}{p(u|H0).p(v|H0)} \qquad (6)$$

where u and v are the training and test i-vectors respectively. Finally score level averaging was conducted for each language,

$$w(l,v) = \frac{\sum_{\forall n} s(u,v)}{n} \qquad (7)$$

where n is the number of utterances in each language and l is the target language.

3.4. NIST 2007 and 2015 LRE Datasets

The NIST 2007 LRE is based on 14 target languages involving conversational speech across speech channels [17]. We used all 14 target languages for training and testing purposes. 10 conversations from each language were randomly selected for development purposes. In keeping with the structure of NIST 2007, our studies were carried out on 30 seconds speech segments for the closed-set condition.

The NIST 2015 LRE features 20 target languages, subdivided into 6 main clusters as shown in Table 1 involving both conversational telephone speech data and broadcast narrowband speech data recorded in various conditions. The language models should only train on these limited and specified training data according to the evaluation plan [18].

Table 1. *NIST 2015 Language clusters with target languages*

Cluster	Target Languages
Arabic	Egyptian, Iraqi, Levantine, Maghrebi, Modern Standard
Chinese	Cantonese, Mandarin, Min, Wu
English	British, General American, Indian
French	West African, Haitian Creole
Slavic	Russian, Polish
Iberian	Caribbean Spanish, European Spanish, Latin American Spanish, Brazilian Portuguese

There are 7020 utterances for target language modelling and as before, we selected 10 random conversation speech segments from each language for development purposes. Unlike NIST 2007, the results obtained were computed over all test segments, having durations of between 3s and 30s according to their primary task.

4. Results

In this section, experiments are carried out on NIST 2007 and NIST 2015 datasets to demonstrate that the use of Eigenfeatures is a better alternative to the shifted delta coefficients in a language identification system. In order to validate the class separability on training data, we calculated the *J*-measure for both NIST 2007 and NIST 2015 LRE data sets. The *J*-measure is the ratio between inter-class scatter to intra-class scatter as shown below,

$$J = trace\ (S_W^{-1}S_B) \qquad (8)$$

where S_W and S_B are the within-class and between class scatter matrix respectively. The larger the value of *J*-measure, the better the discrimination of the classes in the feature space. The results in Table 2 show a better class separability using Eigenfeature than shifted delta coefficient features that leads to a better performance.

Table 2. *J-measure for PLLR and BNF features concatenated with SDC/ EF features*

Features	*J- measure*	
	NIST 2007	NIST 2015
PLLR	10.17	12.03
PLLR_SDC	10.39	12.17
PLLR_EF	**10.45**	**12.25**
BNF	10.76	12.86
BNF_SDC	10.96	13.08
BNF_EF	**10.99**	**13.11**

In our work, all LID systems were compared in terms of average cost performance, (Cavg) and Log likelihood ratio function, (Cllr) provided by the NIST 2007 and NIST 2015 LRE evaluation plans respectively. The results of the experiments using NIST 2007 LRE data set are summarized in Table 3 with the first three rows corresponding to PLLR features and last three corresponding to BNF features. It can be clearly seen that in both cases Eigenfeatures outperformed SDC features having relative improvement of 10.8% in PLLR and 42.3% in BNF.

Table 3. *%Cavg and Cllr performance for the standard PLLR and BNF features compared with concatenated SDC and EF features for NIST 2007 LRE*

System	%C_{avg} (C_{llr})
PLLR	3.88 (0.244)
PLLR_SDC	3.16 (0.208)
PLLR_EF	**2.82 (0.189)**
BNF	1.85 (0.142)
BNF_SDC	1.63 (0.132)
BNF_EF	**0.94 (0.0818)**

To validate the consistency of Eigenfeatures in different databases the above techniques were applied to the NIST 2015 LRE data set. The results presented in Table 4 show a relative improvement of 3.3% on PLLR and 3.9% on BNF in terms of %Cavg using EF compared to SDC features.

Table 4. *%Cavg and Cllr performance for the standard PLLR and BNF features compared with concatenated SDC and EF features for NIST 2015 LRE*

System/ Cluster	%C_{avg} (C_{llr})					
	PLLR	PLLR_SDC	PLLR_EF	BNF	BNF_SDC	BNF_EF
Arabic	29.2 (2.04)	27.7 (1.93)	**27.5** **(2.38)**	27.3 (2.30)	26.9 (2.71)	**25.9** **(2.34)**
English	22.6 (1.59)	18.0 (1.11)	**17.5** **(1.28)**	14.5 (0.97)	12.6 (0.84)	**13.9** **(0.89)**
French	42.5 (5.20)	43.3 (5.81)	**42.0** **(5.89)**	40.6 (5.25)	41.9 (5.88)	**39.0** **(6.37)**
Slavic	9.33 (0.434)	8.47 (0.40)	**8.37** **(0.49)**	6.24 (0.36)	6.09 (0.33)	**5.61** **(0.32)**
Iberian	26.2 (2.19)	26.8 (1.90)	**23.9** **(2.38)**	21.5 (1.87)	21.6 (2.08)	**20.9** **(2.22)**
Chinese	22.4 (1.94)	20.9 (1.63)	**20.9** **(1.92)**	16.7 (1.47)	15.0 (1.34)	**14.3** **(1.35)**
AVERAGE	25.4 (2.23)	24.2 (2.13)	**23.4** **(2.39)**	21.1 (2.04)	20.7 (2.19)	**19.9** **(2.25)**

In order to improve the system performance further, score fusion was conducted. Fusion parameters were estimated and applied to the development data sets using the Focal toolkit [19]. Table 5 shows the results for fusion of different combinations of systems. The results suggest that the fusion of PLLR_EF system with BNF_EF system outperforms any other combinations for the 2007 and 2015 LRE data sets used in our experiments.

Table 5. *Fusion of LID systems and its performances*

Dataset	Fusion of LID Systems	%C_{avg} (C_{llr})
2007 LRE	(PLLR)+(BNF)	1.03 (0.0377)
	(PLLR_SDC)+(BNF_SDC)	0.55 (0.0213)
	(PLLR_EF)+(BNF_EF)	0.45 (0.018)
2015 LRE	(PLLR)+(BNF)	20.2 (0.529)
	(PLLR_SDC)+(BNF_SDC)	19.5 (0.502)
	(PLLR_EF)+(BNF_EF)	18.9 (0.499)

5. Conclusions

In this paper, we proposed a regional covariance based Eigenfeatures extraction which can capture the short time varying information for language identification tasks. As discussed above, the improvement of results implies that Eigenfeatures contain certain scale and direction invariance over the C_R regions in each language. Furthermore, this directional invariance is increased by choosing Eigenvectors with the highest Eigenvalues.

In contrast to SDC calculation, these Eigenfeatures capture both temporal information as well as language specific information and can be applied to both PLLRs and bottleneck features which are the state-of-the-art phonotactic and acoustic features respectively. The results included in this paper illustrates specifically that concatenating standard features with Eigenfeatures will be more beneficial than concatenating them with SDC features in the front-end of language identification systems.

6. References

[1] E. Ambikairajah, H. Li, L. Wang, B. Yin, and V. Sethu, "Language identification: a tutorial," *Circuits and Systems Magazine, IEEE,* vol. 11, pp. 82-108, 2011.

[2] S. B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *Acoustics, Speech and Signal Processing, IEEE Transactions on,* vol. 28, pp. 357-366, 1980.

[3] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *the Journal of the Acoustical Society of America,* vol. 87, pp. 1738-1752, 1990.

[4] M. Diez, A. Varona, M. Penagarikano, L. J. Rodriguez-Fuentes, and G. Bordel, "On the use of phone log-likelihood ratios as features in spoken language recognition," in *Spoken Language Technology Workshop (SLT), 2012 IEEE*, 2012, pp. 274-279.

[5] F. Richardson, D. Reynolds, and N. Dehak, "A Unified Deep Neural Network for Speaker and Language Recognition," *arXiv preprint arXiv:1504.00923,* 2015.

[6] P. A. Torres-Carrasquillo, E. Singer, M. A. Kohler, R. J. Greene, D. A. Reynolds, and J. R. Deller Jr, "Approaches to language identification using Gaussian mixture models and shifted delta cepstral features," in *INTERSPEECH*, 2002.

[7] F. Allen, E. Ambikairajah, and J. Epps, "Language identification using warping and the shifted delta cepstrum," in *Multimedia Signal Processing, 2005 IEEE 7th Workshop on*, 2005, pp. 1-4.

[8] M. Diez, A. Varona, M. Penagarikano, L. J. Rodriguez-Fuentes, and G. Bordel, "New insight into the use of phone log-likelihood ratios as features for language recognition," in *INTERSPEECH*, 2014, pp. 1841-1845.

[9] M. Sahidullah and T. Kinnunen, "Local spectral variability features for speaker verification," *Digital Signal Processing,* vol. 50, pp. 1-11, 2016.

[10] O. Tuzel, F. Porikli, and P. Meer, "Region covariance: A fast descriptor for detection and classification," in *Computer Vision–ECCV 2006*, ed: Springer, 2006, pp. 589-600.

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

[12] J. Pelecanos and S. Sridharan, "Feature warping for robust speaker verification," 2001.

[13] S. Irtza, V. Sethu, P. N. Le, E. Ambikairajah, and H. Li, "Phonemes Frequency Based PLLR Dimensionality Reduction for Language Recognition," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[14] P. Schwarz, "Phoneme recognition based on long temporal context," Ph.D. thesis, Faculty of Information Technology, Brno University of Technology, http://www.fit.vutbr.cz/ ,Brno, Czech Republic, 2008.

[15] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel*, et al.*, "The Kaldi speech recognition toolkit," in *IEEE 2011 workshop on automatic speech recognition and understanding*, 2011.

[16] D. Martınez, O. Plchot, L. Burget, O. Glembek, and P. Matejka, "Language recognition in ivectors space," *Proceedings of Interspeech, Firenze, Italy,* pp. 861-864, 2011.

[17] A. F. Martin and A. N. Le, "NIST 2007 Language Recognition Evaluation," in *Odyssey - The Speaker and Language Recognition Workshop*, 2008.

[18] "The 2015 NIST Language Recognition Evaluation Plan (LRE15)," 2015.

[19] "FoCal, Toolkit for Evaluation, Fusion and Calibration of statistical pattern recognizers http://sites.google.com/site/nikobrummer/focal," 2008.

Secondary Tongue Retraction in Arabic Emphatics: An Acoustic Study

Hamed Altairi, Catherine Watson, Jason Brown

The University of Auckland

halt239@aucklanduni.ac.nz, c.watson@auckland.ac.nz, jason.brown@auckland.ac.nz

Abstract

This study provides an acoustic investigation of the secondary articulation of Arabic emphatic sounds. The study observed seven male Arabic speakers from four dialects to investigate F1 and F2 of /aː/, /iː/ and /uː/ in emphatic and non-emphatic environments in 501 monosyllabic words. A repeated-measures ANOVA was conducted to examine the main and interaction effects of emphasis with consonant, vowel, and position factors. Regardless of the inter-subject variability, the Gaussian classification model shows limited differences among subjects.

Index Terms: Arabic, dialect, emphatics, formants

1. Introduction

In standard Arabic, the emphatic coronal consonants include /ṭ/, /ṣ/, /ḍ/, and /ẓ/ contrasted with /t/, /s/, /d/, and /ð/ respectively. Examples of minimal parts are /ṭaːb/ ‘covered’vs. /taːb/ ‘repented’, and /ṣaːb/ ‘targeted’vs. /saːb/ ‘left’. Most of the Arabic dialects have at least three of these consonants and sometimes they have more than four. These sounds are produced with a primary articulation that involves the front of the tongue and a secondary articulation that involves the back of the tongue. Phonetically, the exact nature of the secondary articulation is subject to some controversy. Previous studies report articulatory correlates such as pharngealization [1, 2, 3] and uvularization [4, 5, 6], and other studies report that these sounds are characterized with both [7]. They are also phonologically problematic as they have been grouped with the guttural class [1, 4, 8] while other studies [9, 10] exclude them from such a class.

Acoustically, most previous studies [4, 6, 7, 9] attest that the acoustic correlates of emphatics are lowering F2 and raising F1 of the adjacent vowels. However, at least one study reported no significant raising of F1 [11]. This study reports on the acoustic properties of the emphatic consonants /ṭ/ and /ṣ/ on the neighboring long vowels /aː/, /iː/, /uː/ using seven native speakers of Arabic. The present study is a part of a wider investigation of the acoustic properties of the emphatics, uvulars and pharyngeals and the articulatory mappings of these sounds using ultrasound.

2. Method

2.1. Subjects

Seven male native speakers of Arabic participated in this study. Due to the lack of speakers from a single Arabic variety at the time of the data collection, the speakers were from four Arabic dialects: 2 Yemeni, 1 Jordanian, 2 Palestinian and 2 Egyptian. All Arabic dialects have the three long vowels /aː/, /iː/, and /uː/, and they also make use of the emphatic stop /ṭ/ and the emaphatic fricative /ṣ/ [12]. The Egyptian speakers were all from Cairo; the Palestinians were from north Palestine, the Jordanian speaker was from Amman, and the Yemeni speakers were from Dhamar. All the speakers were living within New Zealand, and they were recorded in Christchurch. All had been in New Zealand for less than five years and they did not have speech or hearing impairments. The speakers were in their mid-twenties and early thirties.

2.2. Stimuli

The target sounds to be investigated were /ṭ/ and /ṣ/ compared to /t/ and /s/ in monosyllabic words of the form CVːC. The three vowels /aː/, /uː/, /iː/ were either preceded or followed by the emphatic or the nonemphatic consonants. The total number of the words is 24 and each word was repeated three times in the carrier phrase [gaːlu.................marratajn] “they said........twice”. The speakers used their local varieties to read the words and there are slight variations in the first sound of the first phrase and the vowel in the last syllable of the second phrase. For example, the Egyptian and Palestinian speakers pronounce the /g/ as /ʔ/.The vowel in the third syllable of the second phrase is pronounced as /eː/ and /iː/ by the Palestinian and Egyptian speakers, respectively. The total tokens for /aː/, /iː/ and /uː/ were 168, 168, and 165 respectively. Three tokens of /uː/ were not clear enough to be analyzed.

2.3. Procedures

The words were audio recorded simultaneously with ultrasound recording at the University of Canterbury in Christchurch, New Zealand. The audio was collected from the microphone of the SONY camera recording. Due to the interference from ultrasound, the resulting signals were very noisy, so no analysis of high frequency sounds was possible. Fortunately, the acoustic energy of the first two formants for the vowels being investigated was less than 2500 Hz, allowing a formant analysis to be performed. The audio files were stored as .wav files digitized at 44.1 KHz. Praat [13] was used for labelling and segmenting the vowels. To calculate the formant values and to create vowel spaces for the speakers, Emu and the EmuR package [14], and ggplot2 package [15] were employed. The statistical procedures were performed using R [16].

Visual inspection for formant tracks was performed for all of the data. There were errors in formant tracking due to the low amplitude of the high vowels /uː/ and /iː/, making it difficult for the automatic formant tracker to track acoustic energy. Hand correction was applied by redrawing the formant tracks for F1 and F2 based on a closer examination of the spectrogram of the vowels; 38% of the data was corrected. With regard to F3, it was difficult to redraw the formant track due to the high noise in the audio signal; therefore, this formant is excluded from the analysis. Based on the default setting in the formant tracker in Emu [14], formant values were calculated for the midpoint of the vowel identified, which is assumed to represent the steady state of the vowel.

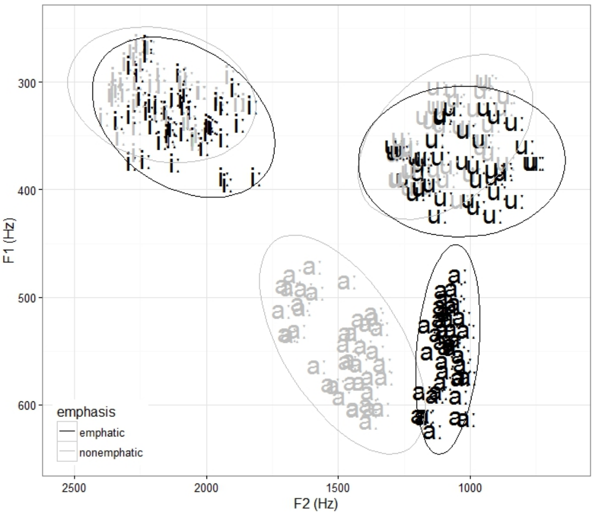


Figure 1: *F1-F2 of the vowels /a:/, /i:/ and /u:/ following the emphatics /ṭ/ and /ṣ/, and non-emphatics /t/ and /s/.*

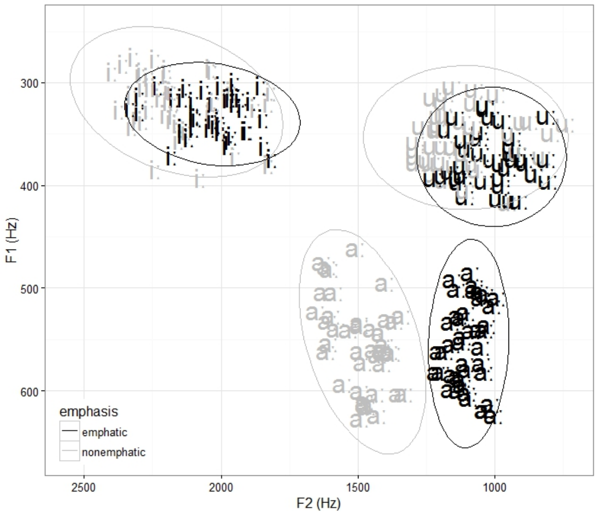


Figure 2: *F1-F2 of the vowels /a:/, /i:/ and /u:/ preceding the emphatics /ṭ/ and /ṣ/, and non-emphatics /t/ and /s/.*

3. Results

Table 1 shows the means for F1 and F2 of the vowels /a:/, /i:/, /u:/ preceded and followed by the emphatics /ṭ/ / and /ṣ/ and non-emphatics /t/ and /s/. The F1 means indicate that it is relatively higher in the emphatic context compared to the non-emphatic environment. F2 of /a:/ before or after the emphatics is substantially lower than F2 of /i:/ and /u:/. The plots in Figures 1 and 2 show the distribution of /a:/, /i:/, and /u:/ in the emphatic and non-emphatic conditions for all speakers in the CV and VC context respectively. The y axis represents F1 and the x axis represents F2. The distribution of the vowels in Figure 1 and Figure 2 exhibits the same behavior, where /a:/ before or after the emphatics has the greatest backing compared to the other vowels.

Table 1: *F1 and F2 means of the vowels for all speakers (E= emphatic, N= non-emphatic).*

	F1		F2	
vowel	E	N	E	N
/a:/	551	548	1090	1484
/i:/	332	315	2059	2166
/u:/	373	352	1021	1100

A series of three-factor ANOVA with repeated measures was performed to test the main and interaction effects of the independent variables (emphasis, vowel, position, and consonant type) on the dependant variables F1 and F2. The factors with 2 degrees of freedom or more need to be tested for the Sphericity assumption, which states the variances of the differences between all possible combinations of groups (levels) are equal. Whenever the assumption of Sphericity was violated, Huynh-Feldt and Greenhouse-Geisser corrections were applied on degrees of freedom and p-value for all within-subject comparisons. The first repeated measures ANOVA was conducted to test the main and interaction effects for the emphasis, vowel and position factors on the dependent factors, F1 and F2. Apart from the highly significant effect of the vowel on F1 (F (2, 12) =190.0, $p[GG] < 0.00$) and F2 (F (2, 12) =215.2, $p[GG] < 0.00$), the result shows that emphasis has a significant effect on both formants (F1: F (1, 6) =13.0, $p < 0.01$, F2: F (1, 6) =187.0, $p < 0.00$). There is also significant two-way interactions by emphasis and vowel relating to F2 (F (2, 12) =27.1, $p[GG]$ <0.00). As for the position factor, there was no significant effect and there were no other interaction effects upon F1 and F2.

The second repeated measures ANOVA was conducted to assess the main and interaction effect of emphasis, vowel and consonant type on the dependent variables F1 and F2. Similar to the previous test, the significant effect of vowel on F1 (F (2,12)=191.5, $p[GG]$ <0.00) and F2 (F (2,12)=222.2, $p[GG] <$ 0.00) was to be expected since there are three different vowels. The main effect of the emphasis factor is significant on raising F1 (F (1, 6) =12.3, p <0.01), but it does not interact significantly with the other factors. Concerning F2, emphasis lowers F2 significantly (F (1, 6) =140.4, p <0.00) and it interacts significantly with the vowel factor (F (2, 12) =26.1, $p[GG]$ <0.00). No other significant main and interaction effects were related to F1 and F2. It is important to note that standard two-way ANOVAs were conducted to determine if the subject variable has effects on F1 and F2 and whether it interacted with emphasis; the results show no significant main or interaction effects upon F1 and F2.

Since emphatics have a significant effect on the two formants, a post-hoc evaluation was carried out to investigate the effects of the emphatic consonants on the two formants of each preceding and following vowel /a:/, /i:/, and /u:/ compared to the formants of these vowels preceding or following the non-emphatic consonants. As illustrated in Table 2, the results indicate that there is a highly significant difference between /a:/ in the emphatic and non-emphatic environment related to lowering F2. Similarly, emphatics cause significant lowering of F2 for the high vowels /i:/ and /u:/ compared to the non-emphatics. F1 is not significantly raised for the low vowel /a:/ when preceded or followed by the emphatic. However, F1 of the high front vowel /i:/ is significantly raised when adjacent to emphatics compared to the non-emphatics. Raising of F1 is also significant in the case of the high back vowel /u:/.

To show there were no dialect effects in the result, an open test Gaussian classification was conducted. On the basis of F1 and F2, the three vowels were categorized as emphatic and

Table 2: *Significant effects of emphatics on raising F1 and lowering F2 of the adjacent vowels (E= emphatic, N= non-emphatic)*

	F1	F2
vowel	E-N	E-N
/aː/	$t(6)$=-0.3 p = .200	$t(6)$=9.1 p < .000
/iː/	$t(6)$=-5.5 p < .004	$t(6)$=4.0 p < .020
/uː/	$t(6)$=-4.2 p < .016	$t(6)$=3.5 p < .036

non-emphatic. Since the experiment employed seven speakers from four dialects, a "round-robin" procedure was performed in which the tokens from a single speaker were used as a test set and the tokens for other speakers were considered as the training set. In Table 3, the Yemeni speaker YEM02 was tested against the remaining 6 speakers. As illustrated below, the diagonals represent the correct classification while other cells show the misclassifications. For example, the emphatic /aː/ was correctly classified with 100% reliability, and there were no misclassifications. The same result obtains with the non-emphatic /aː/. However, the classification rate of /iː/ in the emphatic and non-emphatic environments was 66% and 76% respectively. Finally, the hit-rate was 58% for the emphaic /uː/ whereas it was 91% for the non-emphatic condition. The overall hit rate across all vowel categories is 81% with correct separation. The same procedure was repeated with all of the speakers in turn and Table 4 presents the overall classification. The hit-rate per vowel class is 98%, 70%, and 61% for the emphatic /aː/, /iː/ and /uː/ and 100%, 61% and 71% for the non-emphatic /aː/, /iː/, and /uː/ respectively. This result indicates that regardless of the different speakers, the vowels were 77% classified correctly on the basis of emphasis.

Table 3: *Classification for YEM02 (E= emphatic, N= non-emphatic).*

		/aː/		/iː/		/uː/	
		E	N	E	N	E	N
/aː/	E	12	0	0	0	0	0
	N	0	12	0	0	0	0
/iː/	E	0	0	8	4	0	0
	N	0	0	3	9	0	0
/uː/	E	0	0	0	0	7	5
	N	0	0	0	0	1	11

Table 4: *Classification for all the speakers (E= emphatic, N= non-emphatic).*

		/aː/		/iː/		/uː/	
		E	N	E	N	E	N
/aː/	E	83	01	0	0	0	0
	N	0	84	0	0	0	0
/iː/	E	0	0	59	25	0	0
	N	0	0	32	52	0	0
/uː/	E	0	0	0	0	50	31
	N	0	0	0	0	24	60

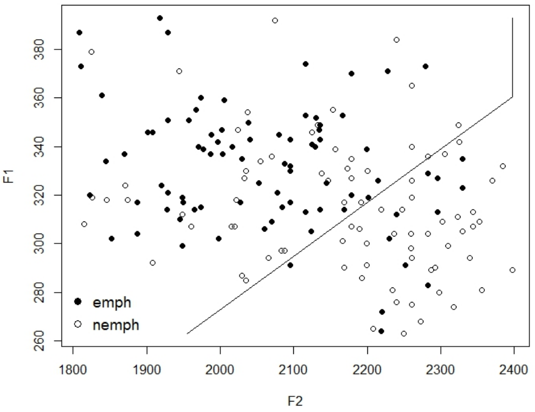


Figure 3: *Scatter plot of F1 and F2 of /iː/ with a discriminant line in the emphatic (emph) and non-emphatic (nmph) environments .*

4. Discussion

The goal of this study was to investigate the midpoint of F1 and F2 for the long vowels (/aː, iː, uː/) preceding and following the emphatics compared to the non-emphatics. The acoustic results of the repeated measures ANOVAs suggest that the emphatic consonants have distinctive coarticulatory effects on the adjacent vowels compared to the non-emphatic consonants across the speakers of the four Arabic dialects (Yemeni, Palestinian, Jordanian and Egyptian). The acoustic study supports earlier studies in that the most salient acoustic feature of the emphatics is lowering F2 [4, 6, 7, 9, 10, 17]. The significance level of the lowering F2 depends on the vowel type. The greatest lowering of F2 occurs with the low vowel /aː/ as shown in Table 1, followed by the high front /iː/ and the high back /uː/. To illustrate, data in Figures 1 and 2 show that the distribution of /aː/ in the emphatic environment is more tightly clustered than the other two vowels, and there is also less overlap. The ANOVAs indicate that the lowering of F2 in the vowels /aː/ and /iː/ can be considered a reliable indicator for the presence of an emphatic consonant. Because a high back vowel /uː/ has a low F2, the impact of emphasis is less.

Although it has been reported that F1 bears no cues for emphasis[11], the present study shows that it is relatively raised in the context of emphatics. Raising F1 is not as salient and systematic as F2. For example, unlike [4] and [6], the repeated measures ANOVA shows that raising F1 is not significant when the low vowel /aː/ is adjacent to the emphatics. With respect to the high vowels /iː/ and /uː/ and similar to [6], the repeated measure and post-hoc tests show that F1 is raised significantly when neighbouring the emphatics. However, when a scatter plot of the /iː/ tokens for all the speakers with a discriminant line as in Figure 3 is observed, the F1 of /iː/ demonstrates less contribution to the classification of the emphatics and non-emphatics. In the case of /uː/, F1 has equal contribution with F2 in categorizing the emphatics and non-emphatics as in Figure 4.

The acoustic findings of this study could be mapped to the behaviour of the tongue dorsum and tongue root. For example, it might be expected that the greater lowering of F2 in /aː/ be mapped to the substantial backing of the tongue dorsum in emphatic contexts. Backing of /iː/ is significant, but limited since

Table of Contents for this manuscript

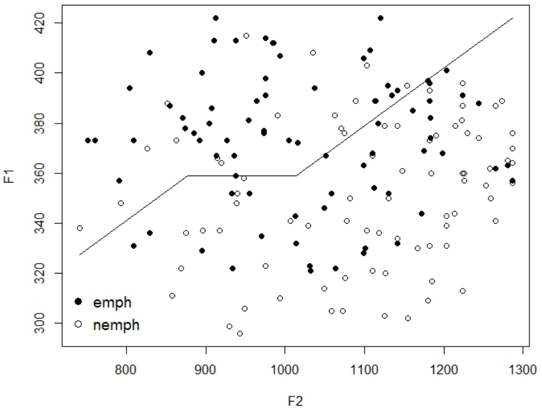


Figure 4: *Scatter plot of F1 and F2 of /u:/ with a discriminant line in the emphatic (emph) and non-emphatic (nmph) environments .*

the high front vowel requires fronting of the tongue body. Such a gesture is antagonistic to the tongue dorsum retraction. With regard to /u:/, the tongue body movement required for this vowel and emphatics are similar as both require tongue dorsum retraction. Hence, the effect of emphatics on this vowel would be less pronounced than in other vowels. Considering the relative rise of F1, it indicates the involvement of tongue root retraction, and such an activity would not be salient and consistent during the articulation of the emphatics. We expect that the tongue root retraction has no role in categorizing the emphatic and non-emphatics in case of the low vowel. In case of the high front vowel, tongue dorsum retraction exhibits more contribution to the classification of this vowel in the emphatic and non-emphatic conditions. Finally, based on the scatter plot, the categorization of the high back vowel in both conditions indicate the involvement of both gestures: tongue dorsum retraction and tongue root retraction.

A recent study [6] reports emphatic stops have a greater effect on adjacent vowels compared to the emphatic fricatives. However, the results of this study show no significant difference between the long vowels when preceded or followed by the emphatic stop and emphatic fricative. While [6] shows that the emphatics in word-final position have a greater effect on the vowels than the emphatics word-initially, the results of this study show that the position of the emphatics initially or finally in the monosyllabic words has a similar effect on adjacent vowels. This acoustic study partially supports the suggestions of [7] that the vowels in the emphatic context undergo two processes: uvularization and pharyngealization. According to [7], only the low vowel undergoes uvularization; however, the current acoustic study shows that the articulation of the vowels /i:/ and /u:/ in the emphatic environment involves consistent tongue dorsum retraction. With respect to pharngealization, [7] reports that only short vowels are pharyngealized, but the current study suggests that long vowels undergo pharyngealization when adjacent to the emphatics.

5. Conclusion

Emphatics are characterized with acoustic effects that are different from the non-emphatics. Lowering F2 and raising F1 are the acoustic attributes of these sounds and based on the behaviour of these two formants, the acoustic results suggest that these sounds are articulated with the tongue dorsum and tongue root retraction. These sounds are categorized well by tongue dorsum retraction represented by the consistent lowering of F2 for all the adjacent vowels. Although F3 may provide valuable information about the articulation of the emphatics, it was not investigated in this study. While [18] reports that F3 is not affected by emphasis, [6] suggests that a raised F3 in the emphatic context indicates that these sounds are uvularized. The third formant of the vowels preceded by the emphatics, uvular and pharyngeal consonants will be discussed in future research. The acoustic results will be then mapped to the tongue dorsum/root retraction of these sound groups using ultrasound.

6. References

[1] S. Davis, “Emphasis spread in Arabic and grounded phonology,” *Linguistic Inquiry*, vol. 26, no. 3, pp. 465–498, 1995.

[2] S. AL-Ani, *Arabic phonology; An acoustical and physiological investigation.* The Hague: Mouton, 1970.

[3] F. Al-Tamimi, F. Alzoubi, and R. Tarawnah, “A videofluoroscopic study of the emphatic consonants in Jordanian Arabic,” *Folia Phoniatrica et Logopaedica*, vol. 61, no. 4, pp. 247–253, 2009.

[4] B. Zawaydeh, “The phonetics and phonology of gutturals in Arabic,” Ph.D. dissertation, Indiana University, 1999.

[5] J. J. McCarthy, “The phonetics and phonology of semitic pharyngeals,” in *Phonological Structure and Phonetic Form*, P. A. Keating, Ed. Cambridge University Press, 1994, pp. 191–233, cambridge Books Online. [Online]. Available: http://dx.doi.org/10.1017/CBO9780511659461.012

[6] A. Jongman, W. Herd, M. Al-Masri, J. Sereno, and S. Combest, “Acoustics and perception of emphasis in Urban Jordanian Arabic,” *Journal of Phonetics*, vol. 39, no. 1, pp. 85–95, 2011.

[7] K. N. Shahin, “Accessing pharyngeal place in Palestinian Arabic,” *Amsterdam Studies in the Theory and History of Linguistic Science Series 4*, pp. 131–150, 1996.

[8] S. Rose, “Variable laryngeals and vowel lowering,” *Phonology*, vol. 13, no. 01, pp. 73–117, 1996.

[9] M. S. Bin-Muqbil, “Phonetic and phonological aspects of Arabic emphatics and gutturals,” Ph.D. dissertation, University of Wisconsin-Madison, 2006.

[10] S. A. M. Shar, “Arabic emphatics and gutturals: A phonetic and phonological study,” Ph.D. dissertation, University of Queensland, 2012.

[11] C. Elizabeth, “A phonetic and phonological study of Arabic emphasis,” Ph.D. dissertation, Cornell University, 1983.

[12] J. C. Watson, *The phonology and morphology of Arabic*. OUP Oxford, 2007.

[13] P. Boersma and D. Weenink, *Praat: Doing phonetics by computer[Computer program]*, 2013, version 5.3.56.

[14] R. Winkelmann, K. Jaensch, S. Cassidy, and J. Harrington, *emuR: Main Package of the EMU Speech Database Management System*, 2016, r package version 0.1.8.

[15] H. Wickham, *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2009. [Online]. Available: http://ggplot2.org

[16] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2014. [Online]. Available: https://www.R-project.org

[17] S. Ghazeli, “The phonetics and phonology of gutturals in Arabic,” Ph.D. dissertation, University of Texas at Austin, 1977.

[18] A. Giannini and M. Pettorino, “The emphatic consonants in Arabic (speech laboratory report),” *Naples: Istituto Universitario Orientale*, 1982.

Sound Change or Experimental Artifact?: A study on the impact of data preparation on measuring sound change.

Catherine I. Watson and Zoe E. Evans

University of Auckland, University of Otago

c.watson@auckland.ac.nz, zoe.evans@otago.ac.nz

Abstract

Forced alignment systems are commonly used to process large amounts of data in socio-phonetic studies. We compare how two of these systems perform against manual segmentation and phonetic labelling in a database of New Zealand English. The results indicate predictable variations in terms of the relative sizes of the vowel spaces, but also suggest the need to be cautious in interpreting small phonetic variations, as these may be the result of the method used to segment and label the data.

Index Terms: sound change, forced alignment, artifact, vowel

1. Introduction

Socio-phonetic studies have been employed over many years to understand the factors that result in speech variability, such as region, age, gender, culture, language exposure, and so forth. The observed variations in these studies can often be small, but significant [1]. For sound change studies it is important to ensure that these small changes are indeed indicative of actual change. To improve the robustness of these findings, large amounts of data are gathered, but the task of manually transcribing, segmenting, and labelling these data sets is a daunting one. Consequently, it has become more common in recent years to involve automated systems in the processing of data from sizeable speech corpora [2-4]. There are many automated processes available to speech researchers, performing a range of tasks from audio dictation (e.g.[5]), to forced alignment and phonetic transcription (e.g. FAVE [6], MAUS [7], DARLA[5], LaBB-CAT [8]), through to formant extraction [9,10].

In this study, we focus on the forced alignment systems that are used to automatically segment and label the speech sounds in an audio file. In order to successfully identify phonemes and their boundaries, these systems use appropriate acoustic models and pronunciation dictionaries. Here, we compare two systems: FAVE-align, and MAUS. The Forced Alignment and Vowel Extraction (FAVE) suite provides two services, one of which is an aligner (FAVE-align) which uses a U.S. English (USE) acoustic model and USE pronunciation dictionary to automatically generate a Praat textgrid containing a word and phonetic tier. Although FAVE has been developed for, and arguably performs well for analyses of USE, it has been gaining traction outside the States for use in British English [2, 11], New Zealand English (NZE) [12], Australian English (AE) [13] and Bequia [21]. Since USE acoustic models may not perform as accurately for non-US Englishes, an alternative aligner was selected for comparison. The Munich AUtomatic Segmentation (MAUS) aligner [7] is an online service of the Bavarian Archive for Speech Signals. MAUS offers acoustic models and lexicons for a range of English dialects (and other languages), and therefore may provide more accurate phonetic segmentation and labelling for non-US Englishes.

While these automated systems vastly speed up the processing of speech data, they may not always match the results provided by a more manual analysis [14]. This distinction is important as some recent studies have used these different methodologies to draw comparisons between different data sets [15]. If automatic and manual alignments do not provide comparable data, this may result in the observation of spurious phonetic variation. The importance of using similar methodologies to compare and contrast results from different data sets cannot be overstated. In a recent comparison of two AE corpora [15], substantial differences were discovered between vowel productions. Given that the speakers in the corpora were different ages, were from different locations, and that the vowel formants were extracted using different techniques, it is difficult to draw any helpful conclusions about the observed variability. Similarly, in another analysis of AE [13], it was found that using a forced aligner to segment and label the data resulted in errors that may have influenced the variability of vowel measurements, particularly for vowel durations.

As more and more socio-phonetic studies rely on automated systems for processing large speech corpora, it is important to establish that the data generated by these systems does not differ substantially from the data generated by hand labelling. In manual segmentation and labelling, researchers carefully select tokens that have minimal phonetic reduction (i.e. from lexically stressed syllables of pitch-accented words), and from environments with minimal coarticulatory effects. Given this, and the fact that forced-alignment systems are not yet capable of identifying pitch-accented words, we would expect that a phonetic analysis using forced alignment would show greater reduction along all dimensions compared with an analysis based on hand-labelled data. If forced alignment systems are to be relied on, we would also hope to see the same patterns of phonetic change, albeit within a reduced vowel space, as are observed in manually labelled data. To this end, in the current study we provide three separate analyses of a single data set. This data was originally presented in [16], and described vowel change over time in three speakers of NZE. The data was manually segmented and labelled. In the present study, we will re-present those results alongside results from the same data set where the segmentation and labelling was carried out automatically by the FAVE aligner, and MAUS. Our aim is to establish to what degree the use of automated aligners and labellers may impact on results.

2. Method

2.1. Database

The speakers were all prominent New Zealand men (5 in total) for which there were multiple radio recordings available [16]. Most of the recordings were interviews. In this study we included 2 more speakers and for each speaker looked at recordings made mainly in the 1950s (henceforth called the 50s set) and recordings made mainly in the 1980s (henceforth called the 80s set). The recordings were digitized to a 16 bit number, at a sampling rate of 20 kHz, and stored as WAV files. The speaker and recording details are in Table 1. Only monophthongs were used in the current analysis.

Table 1: The Speakers' date of birth and recordings.

	A	B	C	D	E
Date of Birth	1919	1901	1916	1935	1935
Recordings:19xx	54,92	54,82	55,85	60,86	60,84

2.2. Hand-Labelled Data Preparation

All the data was hand labelled at the phonetic level using EMU labeller [10]. Only the vowels from prosodically accented words were selected for analysis. The formant were automatically tracked in ESPS/WAVES+ (12th order LPC analysis, cosine window, 49 ms frame size, 5 ms frame shift). All formant data was checked and corrections were made by hand if necessary. The first formant (F1) and second formant (F2) values at the vowel target were obtained. The vowel targets were identified by hand. See [17] for the guidelines and [16] for the phonetic contexts. This resulted in 3127 monophthongs for the hand-labelled corpus (henceforth called the HAND data).

2.3. System-Labelled Data Preparation

Transcripts were generated for each of the segmented WAV files. The transcripts and their associated WAVs were then passed to two automated segmentation and labelling systems: MAUS [7], and FAVE (Forced Alignment and Vowel Extraction) [6]. The MAUS and FAVE systems returned PRAAT textgrids containing phonetic tiers. The vowel formants were calculated using the same package as the HAND data, with the same analysis settings. The first and second formant values used in this study were extracted from the vowel midpoint but no checking was performed

2.3.1. MAUS and FAVE Data Preparation

The MAUS alignment system allows the user to select from a number of English dialects. In this study, the NZE option was selected, and PRAAT textgrids were generated. Three levels were identified in each textgrid; word, phoneme, and phonetic. At the phonetic level some vowels were identified as schwa. This is determined both by the lexicon and acoustic segmentation process. This resulted in 11602 monophthongs for the MAUS labelled corpus (hence forth called MAUS data).

The textgrids generated by FAVE align contained two tiers: a word and phonetic level. The FAVE phonetic transcription is based on US English pronunciation, which may result in phonetic confusions in non USE dialects. All words in the transcripts were either already within FAVE's internal lexicon, or needed to be added to a separate dictionary. FAVE-align assigns all vowels primary, secondary, or no lexical stress. Only vowel tokens marked with primary lexical stress were included in this study. This resulted in 9612 for the FAVE labelled corpus (henceforth called FAVE data).

2.4. Vowel Space Size and Sound Change Measurements

Analysis of the formant data was in R [18]. Two measures were used to compare the data preparation approaches: one for the vowel space size and one for the movement of vowels. To calculate the vowel space size for each speaker we estimated its width by finding the difference in F2 (in Bark) between FLEECE, and THOUGHT centroids (see figure 1), and estimated its height via the difference in F1 (in Bark) between FLEECE and START centroids (see figure 1).

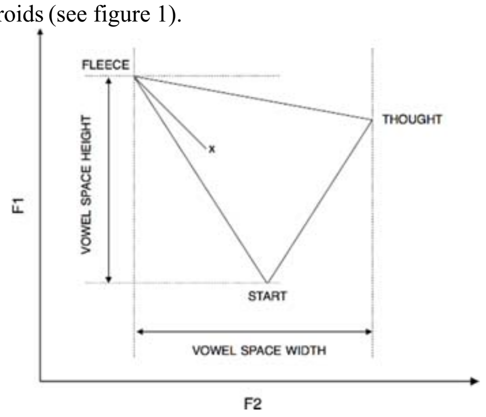


Figure 1 Stylised vowel space with point vowels

To measure the extent of the vowel shifts we used the Vowel Space Movement (VSM) measure [19]. Four vowel shifts are investigated: DRESS raising, KIT falling, KIT retracting, and GOOSE fronting. All have occurred over the timespan of the data (see e.g. [16,19]). The movements the vowel space are be measured by calculating

$$VSM(x)\frac{EU(i\text{:},x)}{EU(i\text{:},pointvowel)} \qquad (1),$$

where x is the vowel of interest, , $EU(i\text{:},x)$ is the Euclidean distance between the FLEECE vowel and x in the F1/F2 vowel space (see Figure 1), *pointvowel* is either START or THOUGHT, depending on the direction of the movement, and $EU(i\text{:},pointvowel)$ is the Euclidean distance between FLEECE and *pointvowel* (see Figure 1). For movements up at the front of the vowel space the point vowel was START, and the x was DRESS or KIT. For movements along the top of the vowel space the point vowel was THOUGHT and x was either KIT or GOOSE. The Euclidean distances are calculated in Bark. The VSM results in a value between 0 and 1. The smaller the value, the closer the vowel in question (x) is to the FLEECE vowel.

3. Results

3.1. Size of the Vowel Space

Figure 2 (placed at end of the paper) contains three F1/F2 plots which have the centroids of each of the NZE monophthongs for the recordings from the 50s set and 80s set from Speaker A for each of the three different data preparation approaches. The 50s data is in the dark hue, and the 80s data is in the light hue. It can be seen the vowel spaces for all three approaches looks reasonably similar. But, as expected, the vowel space from the HAND data has a greater range in both F1 and F2 than that from the MAUS and FAVE data. Vowel shifts for each speaker between the 50s and 80s sets are apparent for all three approaches, however the extent of these changes potentially differs, for example the differences in the TRAP, NURSE and GOOSE centroids. For the FAVE data there were issues with LOT, START, and THOUGHT labels, due to differences between NZE and USE. We were able to correct some of these checking the labels of the corresponding vowels in the MAUS data.

To investigate the difference in vowel space size for all speakers we did two repeated measures ANOVA with vowel space width and height being the dependent variables, respectively and data preparation method, and year of recording being the within-subject factors. Sphericity of the within-subject factors was detected and Greenhouse–Geisser

Table of Contents for this manuscript

adjustments were made to the p values. For both the height and width measures there were significant differences due to data preparation approaches (Width: F(2,8)=44.9 p_{GG}<0.01, Height: F(2,8)=40.4 p_{GG}<0.01) but nothing else. Since the year of recordings (hence age of the speaker) was not a significant factor for vowel space size combined the data from the two different years. A paired t-test (corrected for multiple comparisons) showed that the width and height from the HAND data was significantly greater at the p<0.01 level than that from the MAUS data (Width: t(9)=9.5, Height: t(9)=4.9), and the FAVE data (Width: t(9)=8.7, Height: t(9)=9.5). There were no significant differences in the dimensions for the MAUS and FAVE data. Table 2 is the mean width and height of the vowel space for the 5 different speakers. The vowel space size for the HAND data is greater than for the MAUS and FAVE. Next we investigated whether the extent of the vowel shift is the same in all three approaches. We used the vowel space measure (VSM) to establish whether the movement of the KIT, DRESS and GOOSE vowels is the same, regardless of the data preparation approach.

Table 2.The width and height of the vowel space in Bark by Speaker the three data preparation approaches.

	A	B	C	D	E
	(Bark)	(Bark)	(Bark)	(Bark)	(Bark)
HAND	6.0, 4.0	6.9, 4.0	5.9, 3.3	5.7, 2.3	4.9, 3.2
MAUS	4.7, 3.3	5.1, 3.2	4.3, 2.5	4.1, 2.0	4.1, 2.3
FAVE	4.7, 3.1	5.0, 3.2	4.6, 2.4	4.2, 1.3	4.4, 1.9

3.2. Measuring Vowel Movements

In presenting these results we first focus on the HAND data. Figure 3 shows the movement of DRESS over time, relative to the front of the vowel space. Looking at the HAND data, for all speakers except D, the VSM value for the 80s data is less than for the 50s data, indicating it is closer to the FLEECE vowel and therefore demonstrates the DRESS raising in this diachronic data. No DRESS raising was noted for Speaker D, but the low VSM 1950s value for the HAND data show his DRESS was already very raised. Figure 4 shows the movement of KIT over time relative to the top of the vowel space. For speakers B,D, and E the VSM values for the 80s data is greater than that for the 50s data, which means it is further away from FLEECE, i.e. retracted. Figure 5 shows the movement of GOOSE over time, relative to the top of the vowel space. With the HAND data it can be seen that for Speakers A and C the VSM values for the 80s data are less than that for the 50s data, indicating it is closer to FLEECE, and therefore demonstrating the well-known GOOSE fronting. Due to lack of space the plot for KIT lowering is not included.

When looking at the VSM values from the MAUS and FAVE approaches, we found the VSMs for the three different methods are all highly and significant correlated, see Table 3. However for any speaker-vowel combination the VSM measures for the vowel movements are rarely similar for the three data approaches and the magnitude of the change between the 50s and 80s data differs across the three methodologies, regardless of speaker.

Figure 6 gives the vowel shift results for each speaker, where red hue indicates the VSM for the 80s data is less than the 50s data, and a green hue indicates it is greater than. The intensity of the colour indicates the magnitude of the difference of the VSM between the 50s and 80s. The greater the difference the darker the hue. It would be expected that the cells for the DRESS rising and GOOSE fronting would all be a reddish hue (VSM being closer to 0 for the 80s data), and the cells for KIT falling and retracting to have a greenish hue (VSM being closer to 1 for the 80s data). There are 20 sound changes investigated (5 speakers X 4 vowel movements). The expected changes happened 12/20 possible times for the HAND and FAVE data, and 15/20 for the MAUS data. In addition the three measures were in complete agreement only 10/20 times (as indicated by the similar hue in all three cells for the speaker), and there were a further 5 where either FAVE or MAUS were in the same direction as the HAND data. However, even when there is agreement in the direction of the vowel change between the three measures, the varying intensity of the cell colour indicates that the three methods rarely agree on the *extent* of the sound change.

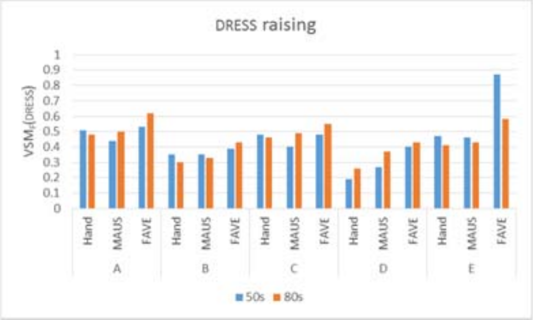


Figure 3: The VSM of DRESS raising in the front of the vowel space for the three different approaches.

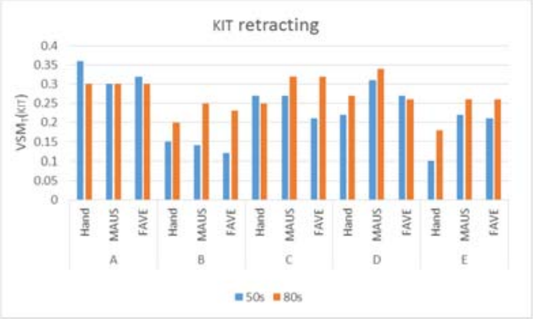


Figure 4 The VSM of KIT retracting along top of the vowel space for the three different approaches.

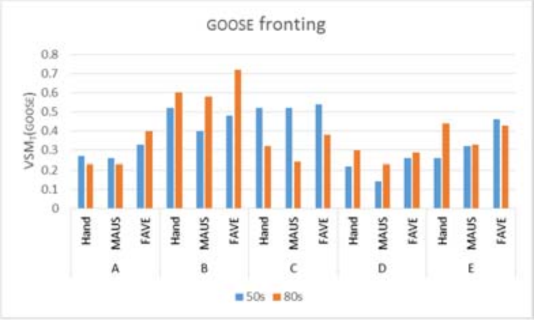


Figure 5: The VSM of GOOSE fronting along the top of the vowel space for the three different approaches.

4. Discussion and Conclusions

In this study we have re-investigated the vowel shift observed in diachronic data of 3 New Zealand English speakers, and included 2 more speakers. We compared three different data preparation approaches: hand labelling (which was used in [16]), labelling via MAUS, and labelling via FAVE.

The three data preparation methods took data from exactly the same set of recordings. However since only vowel tokens with sentence stress (and word stress) were selected for the HAND data, this labelling approach had considerably fewer tokens for analysis compared to the other two. In those two cases only unstressed vowels were excluded from analysis. This yielded 3-4 times more tokens than in the HAND instance, but also resulted in significantly reduced vowel spaces. Without some form of natural language processing it is not possible to automatically select such tokens from MAUS and FAVE. We used the VSM to measure the extent of the vowel shifts. Our studies found that all three methods give highly significant correlated VSMs. However there is differences in the extent of the measured vowel shift. The three data approaches are only in complete agreement of the direction of movement ~50% of the time, and a further 25% of the time either FAVE or MAUS data agrees with the HAND data.

Table 4: Correlations of the VSM for data preparation approaches

	HAND VS MAUS: R,P	HAND VS FAVE: R,P	MAUS VS FAVE: R,P
50s	0.84,p<<0.01	0.72,p<<0.01	0.80,p<<0.01
80s	0.79,p<<0.01	0.89,p<<0.01	0.78,p<<0.01

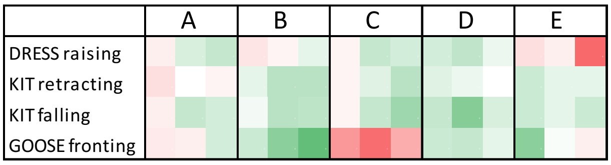


Figure 6 A heat plot of the vowel shift results for the HAND, MAUS, and FAVE data: red indicates 80s VSM < 50s VSM, and green indicates 80s VSM > 50s VSM.

This study has focused mainly on the comparison between the three data labelling approaches. However it is worthwhile to note that in terms of measuring sound change, perhaps this diachronic data is not the best data set for this. In contrast to other studies (eg. [20]) the differences in the speech of the 5 individuals over time is not overwhelming. In conclusion whilst there is a general agreement across the three methods in the direction of the sound change but there is little agreement on the extent of the sound change.

5. Acknowledgements

Thanks to Sallyanne Palethorpe for help in data preparation, thanks to Miriam Meyerhoff for feedback on the paper, and thanks to Aldo Cortesi for his python assistance.

6. References

1. Stuart-Smith, J., Jose, B., Rathcke, T., Macdonald, R. &Lawson, E., ‘Changing sounds in a changing city: An acoustic phonetic investigation of real-time change across a century of Glaswegian’, In E. Moore and C. Montgomery (eds), A Sense of Place, Cambridge: Cambridge University Press, (in press)
2. Cham, B.Q., “Sixty years of speech: A study of language change in adulthood”, Lifespans & Styles 2(1): 17-26, 2016
3. Stanford, J., Severance, N. & Baclawski, K.,”Multiple vectors of unidirectional dialect change in eastern New England”. Language Variation and Change 26:103-140, 2014.
4. Labov, W., Rosenfelder, I., & Fruehwald, J..”One hundred years of sound change in Philadelphia: Linear incrementation, reversal, and reanalysis”. Language, 89(1), 30-65 , 2013.
5. Reddy, S. and Stanford, J., “Toward completely automated vowel extraction: Introducing DARLA.” Linguistics Vanguard. 2015
6. Rosenfelder, I., Fruehwald J., Evanini K, & Y. Jiahong, FAVE (forced alignment and vowel extraction) program suite. http://fave.ling.upenn.edu, 2011
7. Kisler, T, Schiel, F. & Sloetjes, H., “Signal Processing via web services: the use case WebMAUS” in Proceedings Digital Humanities, Hamburg Germany, 30-34, 2012
8. Fromont, R. & Hay, J. “LaBB-CAT: An annotation store.”: Australasian Language Technology Workshop (ALTA), 4-6 Dec 2012. In Proceedings 10:: 113-117, Dunedin, 2012.
9. Boersma P. & Weenink D,: Praat: doing phonetics by computer. Version 5.3.51, http://www.praat.org/ ,2013.
10. Cassidy, S. and J. Harrington, “Multi-level annotation in the Emu speech database management system”, Speech Comm., 33, 61-77, 2001
11. MacKenzie, L., & Turton, D.. “Crossing the pond: Extending automatic alignment techniques to British English dialect data.”New Ways of Analyzing Variation, 42, 2013.
12. Hazenburg, E. “Gendering vowels: Sociophonetics and community affiliation”. NZ Ling. Soc. Conf., Dec, Dunedin. 2015
13. Nguyen, N. & Shaw, J.A., “Why the SQUARE vowel is the most variable in Sydney”, in Proc. 15th Aust. Int. SST 36-39, 2014
14. Severance, N., Evanini, K. & Dinkin, A., “Examining the reliability of automated vowel analyses using FAVE.” Presented at NoWPhon 2, Eugene, Ore, 2016
15. Elvin, J. & Escudero, P., “Comparing acoustic analyses of Australian English vowels from Sydney: Cox (2006) versus AusTalk” in Proc. of the 15th Aust. Int. SST Conf., 44-48, 2014
16. Watson, C., Palethorpe, S., & Harrington, J. (2004). Capturing the vowel change In New Zealand English over a thirty year period via a diachronic study. In Proc. 10th Aust. Int. SST, 201-206, 2004.
17. C. I. Watson, J. Harrington & Z. Evans "An acoustic comparison between New Zealand, and Australian English vowels", Australian Journal of Linguistics, 18 (2), 185-207, 1998
18. R Core Team R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/, 2015.
19. Watson, C. I., Maclagan, M., King, J., Harlow, R., & Keegan, P. J. “Sound change in Māori and the Influence of New Zealand English”. Journal of the International Phonetics Association DOI 10.1017/S0025100316000025, 2016.
20. Harrington, J., Palethorpe, S., & Watson, C. I., “Does the Queen speak the Queen's English?, Nature, 408(6815), 927-928, 2000
21. Walker, J. A & M. Meyerhoff “Mergers of the Caribbean: Low-back vowels in Bequia” Paper presented at Change and Variation in Canada 8. Queen’s University, Ontario., 2014.

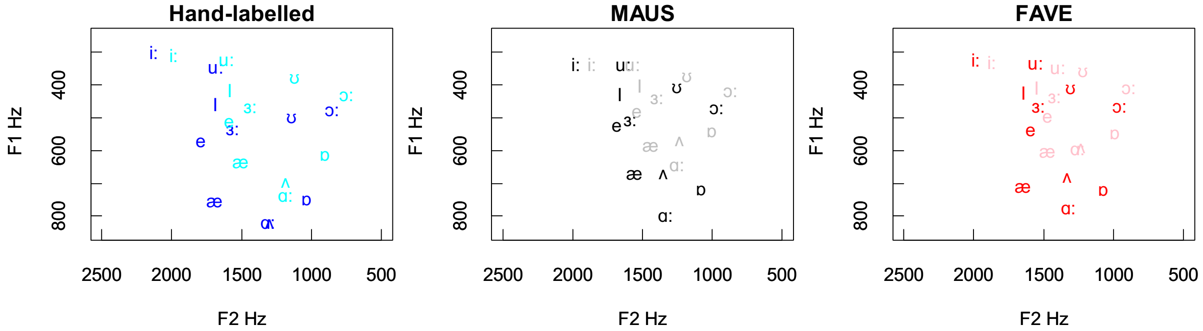


Figure 2: The vowel space of Speaker A from the 50s data (dark) and 80s data (light) for the three different data approaches

Temporal correlates of Lopit singleton and geminate glides

Rosey Billington
University of Melbourne
rbil@unimelb.edu.au

Abstract

Length contrasts among glides are typologically uncommon, and argued to be crosslinguistically dispreferred. Nevertheless, such contrasts are attested in various languages around the world, though phonetic explorations remain very limited. This paper presents selected findings pertaining to glides in Lopit, an Eastern Nilotic language for which a length distinction has been proposed for palatal and labiovelar glides. Duration values are compared for intervocalic tokens of putative geminates /wː/ and /jː/ and singletons /w/ and /j/, and for vowels preceding them, and indicate that in Lopit, duration is a major correlate distinguishing glides at the same place of articulation.

Index Terms: geminate, length, glide, duration, Nilotic

1. Introduction

1.1. Glides and geminate typology

Crosslinguistic surveys of consonant gemination[1] note enormous diversity in quantity contrasts in the world's languages, but certain patterns have also emerged, including that less sonorous segments such as stops, and particularly voiceless stops, are among the most preferred consonants for contrasts based on length. Glides such as [w] and [j] are among the least likely segments to be found geminated [1]. A relationship between sonorancy and the markedness of geminacy has been proposed, with suggestions that the spectral continuity of glides in relation to adjacent vowels likely hinders the perception of differences in constriction duration [2]. It has also been suggested that in production, phonemic differences in duration may be less robust for glides than for other manners of articulation [3]. However, very little acoustic phonetic research has taken place to examine the characteristics of geminate compared to singleton glides which, despite their rarity (and in some cases presumed impossibility), are described for various Indo-European, Finno-Ugric, Dravidian, Austronesian, Oto-Manguean, Afro-Asiatic, Niger-Congo, and Nilo-Saharan languages [1].

1.2. Geminate consonants in Lopit

Existing observations for Lopit, an Eastern Nilotic (Nilo-Saharan) language of South Sudan and its diaspora, all include proposals that some consonants of the same place, manner and voicing have a distinction involving, to some extent, length differences [5] [6] [7]. Data collected in the wider documentation project this study is part of similarly point to such a contrast which, based on impressions of length, has been referred to as a contrast between geminates, such as /tː, dː, nː, lː, r, wː, jː/, and singletons, such as /t, d, n, l, ɾ, w, j/. With the exception of /tː, t/, alveolar length contrasts in Lopit have a low functional load, but for the glides, both putative geminate /wː, jː/ and singleton /w, j/ are found reasonably often, and the contrast is present word-initially as well as word-medially. Lopit is one of several Eastern Nilotic languages for which glide contrasts variously described as involving length or strength are proposed, though none have been the subject of phonetic investigation. For at least the long/strong labiovelar glides in Eastern Nilotic, an origin in stop-glide sequences has been tentatively suggested [5].

1.3. Phonetic correlates of gemination

Across languages, the most consistent phonetic correlate of consonantal length contrasts is constriction duration, which, given that most studies investigate obstruent length, generally refers to the period of complete closure between articulators. Overviews show that geminate consonants reliably have higher duration values than singletons, and furthermore, that vowels preceding geminate consonants often have lower duration values [8] [9] [10], though in some cases they may be no different, or may be longer [11] [10]. For stops, geminates are generally 1.5-3 times longer than singletons [12], and while very little duration data is available crosslinguistically for glides, findings for geminate glides in languages such as Buginese, Madurese, Lebanese Arabic, Egyptian Arabic, Persian, and Guinaang Bontok seem to accord with this pattern, being approximately 1.4-2.6 times longer than singletons [13] [14] [15] [3] [2]. Where reported, differences on the basis of other acoustic measures are also often found for geminates compared to singletons, indicating that additional language-specific cues may support length contrasts.

2. Research aim

Given that a contrast between two types of labiovelar glide and two types of palatal glide has been proposed for Lopit, and that impressionistic observations suggest that glides at the same place of articulation may be either short or long, the primary aim of this study is to establish whether the temporal characteristics of putative /wː, jː/ and /w, j/ are indicative of a contrast involving length. The focus of this paper is on the constriction duration of intervocalic glides, and the duration of vowels preceding them. Amplitude and formant measures are also considered in ongoing work, but are not reported here.

3. Method

3.1. Participants

The participants in this study were five adult speakers of the Dorik dialect of Lopit: three men (AL, DA, VH) and two women (EA, JT). They are part of a small Lopit community in Melbourne, Australia, whose members arrived in Australia from 2000 onwards. All are multilingual, as is common in South Sudan; additional languages include English and Juba Arabic, but Lopit is the primary language used at home.

[1] Here referring only to contrastive differences in consonant length.

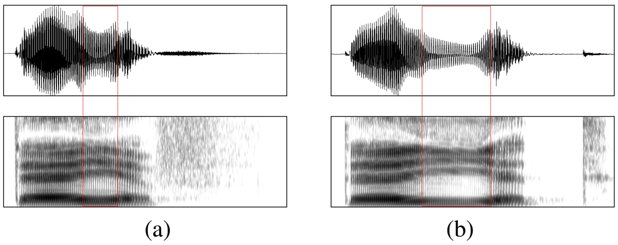

(a) (b)

Figure 1: *Spectrogram and waveform (880 ms) for palatals in (a) [tɛ́jɛ̀f] "IMP.chop" and (b) [tɛ́jːɛ̀t] "IMP.pull".*

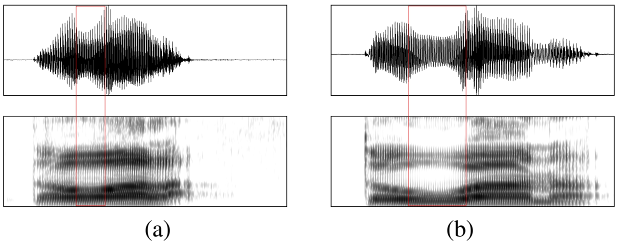

(a) (b)

Figure 2: *Spectrogram and waveform (800 ms) for labiovelars in (a) [tɔ́wálàʔ] "IMP.cough" and (b) [tɔ́wːánàʔ] "IMP.stay".*

3.2. Materials and procedures

A large set of nouns and verbs containing intervocalic examples of /wː/, /jː/, /w/, and /j/ was compiled as stimuli, drawn from the database constructed in the course of the wider project. The words had a range of tonal patterns, and were of mainly 2-3 syllables (with some 4-syllable words); those used in the analysis had glides flanked by non-close vowels /e, ɛ, o, ɔ/ and particularly /a/ [4]. Each word was recorded 5 times in isolation, with occasional additions or omissions, following a spoken English prompt (simultaneously shown on a laptop screen). Data were recorded in a quiet room at a sampling rate of 44.1kHz and 16-bit depth, using a Zoom H4N audio recorder, MixPre-D pre-amp, and AudioTechnica AT892c headset microphone. While the possibility of frication among Lopit geminate glides has been suggested [6], and it was anticipated that this may necessitate the exclusion of some tokens (as they would be unsuitable for the spectral analyses being performed in ongoing work), this was only the case for one hyperarticulated token. Others were excluded for different reasons affecting formant tracking (coughing, creakiness, or formants simply being too weak). The final dataset contained 2384 token representing 91 words (/j/ = 604, /jː/ = 648, /w/ = 572, /wː/ = 560).

3.3. Data processing and analysis

Data were labelled with reference to wideband spectrograms and corresponding waveforms (e.g. Figures 1 and 2) in Praat [16]. Labelling of glides is known to be a challenge. While some previous studies of glides and gemination have used formant cues as the primary criteria for segmentation, from a generous approach taking the period between the end and beginning of steady states of neighbouring vowels [3] to a more conservative approach targeting glide steady states [15], in this study amplitude criteria were taken as the primary indicators of the constriction period of glides. Glides are known to correlate with an overall drop in amplitude, in addition to reduced amplitudes for F2-F4 (particularly F3 for labiovelar glides, and F2 for palatal glides) [17] [18]. These cues are argued to provide the most reliable landmarks for boundaries between glides and vowels, e.g. in automatic segmentation [19], and have been utilised in some studies of singleton and geminate glides [14]. As such, glides in the present data were labelled based on marked changes in overall and upper formant amplitude relative to adjacent vowels. Labelled data were imported into the Emu Speech Database System [20], and duration measures were queried and plotted in the R software environment [21] and statistically tested with Linear Mixed-Effects Models (discussed further below) and post-hoc Tukey HSD tests, using the `lme4` package [22].

4. Results

4.1. Duration of glides

Duration values for /wː/, /jː/, /w/, and /j/ were compared, and, as shown in Figure 3, clear and consistent differences between /wː/, /jː/ and /w/, /j/ can be observed across the five participants, with higher duration values apparent for geminates /wː/, /jː/ compared to singletons /w/, /j/. To investigate the main effect of glide category on glide duration, the data were submitted to a mixed-effects model with glide category as a fixed effect and participant and word as random effects, following comparisons between different models. A likelihood ratio test shows the effect of glide category is significant (χ^2(3, N=2384)= 250.87, p<0.001). Post-hoc tests reveal that duration differences between geminate glides and their singleton counterparts are significant; geminate labiovelar /wː/ is an estimated 65 ± 4 ms longer than singleton labiovelar /w/ (p<0.001), and geminate palatal /jː/ is an estimated 74 ± 4 ms longer than singleton palatal /j/ (p<0.001). Geminate labiovelar /wː/ is also 67 ± 4 ms longer than singleton palatal /j/ (p<0.001), and geminate palatal /jː/ is 72 ± 4 ms longer than singleton labiovelar /w/ (p<0.001). There are no significant duration differences between geminate /wː/ and geminate /jː/ (p=0.318), nor between singleton /w/ and singleton /j/ (p=0.945).

For these medial glides in words produced in isolation, duration values are high in general; means for the singletons /w/ and /j/ are 97 ms (sd 24 ms) and 93 ms (sd 22 ms) respectively, and means for geminates /wː/ and /jː/ are 163 ms (sd 33 ms) and 167 ms (sd 34 ms). There is slightly more variation in duration values for geminate glides, some of which is likely related to word length, as shown in Figure 4; a test of the effect of glide length categorised by occurrence in two-syllable compared to 3-syllable words (χ^2(7, N=2251)= 277.91, p<0.001) shows the same significant and substantial duration differences between geminates and singletons, but also shows that geminate glides in two-syllable words (in which they are onsets of the second syllable) are significantly longer than those in three-syllable words (in which they are mostly also onsets of the second syllable, but sometimes of the third) by 29 ± 4 ms for /jː/ (p<0.001) and 15 ± 5 ms for /wː/ (p<0.05). Interestingly, there are no duration differences for the singletons in two-syllable compared to three-syllable words. Overall, geminate glides are 1.77 times longer than singleton glides. Some differences in the ratios of singleton to geminate duration can be observed, as shown in Table 1; geminate labiovelars are 1.71 times longer than singletons, while geminate palatals are 1.82 times longer, and, although the factor of sex did not make any difference to the statistical model, the two female participants (EA and JT) tend towards slightly higher duration values for singleton glides and accordingly have lower singleton to geminate duration ratios.

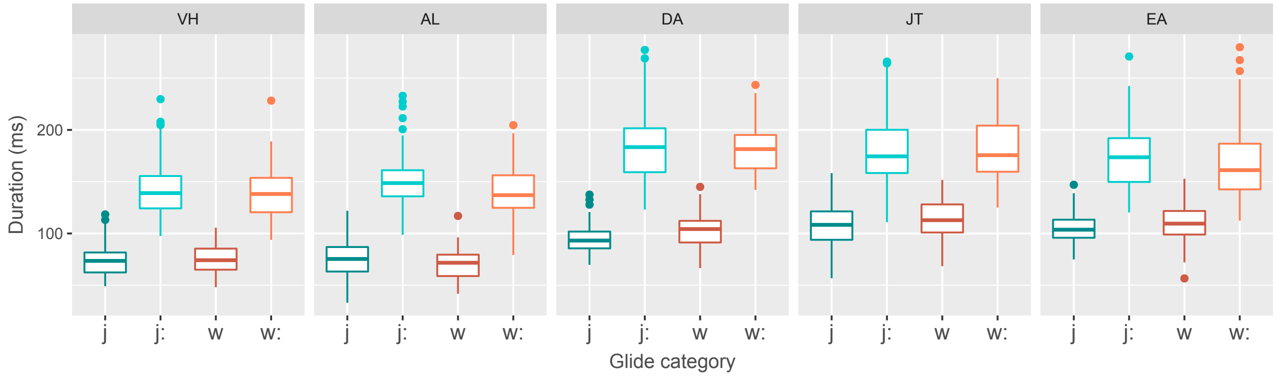


Figure 3: *Duration of proposed singleton and geminate glides in Lopit, for each participant.*

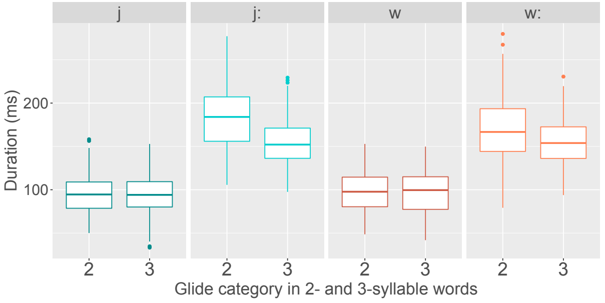


Figure 4: *Duration of singleton and geminate glides in 2-syllable and 3-syllable words, for all participants.*

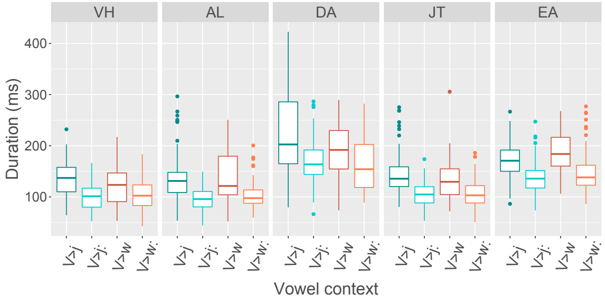


Figure 5: *Duration of vowels preceding singleton and geminate glides in Lopit, for each participant.*

Table 1: *Singleton to geminate duration ratios.*

participant	/w/ : /w:/	/j/ : /j:/	/C/ : /C:/
VH	1 : 1.85	1 : 1.95	1 : 1.90
AL	1 : 1.99	1 : 2.00	1 : 2.00
DA	1 : 1.77	1 : 1.94	1 : 1.85
JT	1 : 1.59	1 : 1.67	1 : 1.63
EA	1 : 1.53	1 : 1.66	1 : 1.59
all	**1 : 1.71**	**1 : 1.82**	**1 : 1.77**

4.2. Duration of vowels preceding glides

Duration values for vowels preceding glides were also compared, and, as seen in Figure 5, there is more variability, but also an apparent tendency towards lower duration values for vowels preceding geminate glides. The effect of glide category was investigated using a mixed-effects model as above, and glide category was found to have a significant effect (χ^2(3, N=2384)= 36.19, p<0.001). In addition, a subset of the data containing only tokens of /a/ was checked, to see if patterns were similar without the possibility of minor duration differences by vowel quality, and the effect of glide category was also significant in this case (χ^2(3, N=1055)= 21.05, p<0.001). Both tests confirm that vowels preceding geminate glides have lower duration values than those preceding singletons; preceding /w:/, vowels are an estimated 28 ± 7 ms shorter than those preceding /w/ (p<0.001), or 37 ± 11 ms shorter when only /a/ tokens are considered (p<0.01), and vowels preceding /j:/ are 34 ± 7 ms shorter than those preceding /j/ (p<0.001), or 34 ± 9 ms shorter for the /a/ subset (p<0.01). In both the test with all vowels and with only open vowels, there are no significant differences in duration between vowels preceding the two geminates /w:/ and /j:/ (p=1.00; p=0.93), nor between vowels preceding the two singletons /w/ and /j/ (p=0.81; p=0.83).

Mean values for vowels preceding /w:/ are 125 ms (sd 42 ms) for all vowels, or 125 ms (sd 36 ms) for only open vowels, and similar for those preceding /j:/, with a mean of 122 ms (sd 41 ms) for all vowels and 119 ms (sd 41 ms) for only open vowels. Means for vowels preceding singleton /w/ are 154 ms (sd 50 ms) for all vowels, or 166 ms (sd 44 ms) for open vowels, and preceding /j/, means are 160 ms (sd 55 ms) for all vowels and 154 ms (sd 54 ms) for open vowels. Vowels preceding singleton palatals are therefore 1.3 times longer than those preceding geminates, and vowels preceding singleton labiovelars are 1.24 times longer; taking both places of articulation together, vowels preceding singleton glides are 1.27 times longer. There are no substantial differences in the duration ratios for individual speakers. While there are hints that segmental and tonal factors may have some influence on preceding vowel durations (for example, high duration values for the small number of vowels with Falling tones), results are for the most part inconsistent except in the case of word length, for which there is a pattern of longer preceding vowels in two-syllable compared to three-syllable words, suggesting a well-attested effect of polysyllabic shortening. When tested statistically (χ^2(7, N=2251)= 118.01, p<0.001), the differences are significant among the singletons, with vowels preceding /j/ being an estimated 47 ± 7 ms longer in two-syllable compared to three-syllable words (p<0.001), and vowels preceding /w/ being an estimated 35 ± 7 ms longer in two-syllable words (p<0.001). Vowels preceding geminate /w:/ are 25 ± 7 ms longer in two-syllable words (p<0.01), but there are no significant duration differences by word length for for vowels preceding /j:/ (p=0.14).

5. Discussion and conclusions

Taken together, these results show that constriction duration is a robust correlate of glide category in pairs of palatal and labiovelar Lopit glides; substantially higher duration values are found for /w:/ and /j:/ compared to /w/ and /j/. Furthermore, geminate glides at both places of articulation are similarly distinct in relation to singletons, suggesting that they are subject to a shared pattern in the consonant inventory. These quantitative findings lend support to impressionistic observations of a length contrast among Lopit glides [5] [6] [7], and bolster arguments that while length contrasts among glides are less common, they have a significant role in the phonologies of some languages [1]. In this data for Lopit, glides /w:/ and /j:/ are overall 1.77 times longer than /w/ and /j/, a ratio in the realm of what is typically observed for geminate compared to singleton consonants [12], and specifically for glide contrasts of this sort in other languages [13] [14] [15] [3] [2]. The trend towards lower duration values for vowels preceding geminates in Lopit also aligns with what is often (though not always) observed for vowels preceding geminate segments in other languages.

This evidence of significant glide duration differences provides a useful starting point for further exploration of consonant length in Lopit as language documentation work continues [23]. Ongoing phonetic research will be well-placed to examine in more detail how the realisation of Lopit glides is mediated by other segmental and prosodic factors, such as word length, given the polysyllabic shortening evident in some of the comparisons discussed here, or vowel quality and tone, which are other aspects of Lopit phonetics and phonology currently receiving close attention [4] [24]. In particular, given that geminate glides are permitted to occur word-intially (albeit infrequently) in contrast with singletons, differences in the production of a typologically less common length contrast in a typologically less common word position for a length contrast will likely offer fruitful insights. However, an obvious next step is to investigate the nature of other proposed length contrasts, all alveolar, at different manners of articulation in Lopit. These other possible length contrasts are found in fewer other Eastern Nilotic languages than the glide contrasts, and may well have a different provenance than the original consonant sequences tentatively proposed for long glides [5].

Finally, it is worth noting that while duration is a clear correlate distinguishing Lopit glides at the same place of articulation, it is unlikely to be the only correlate; additional impressions noted for Lopit [6] and other Eastern Nilotic languages hint at the possibility of articulatory differences, which would not be unexpected for geminate consonants more generally, nor for other consonant types correlating with longer durations. While not addressed here, formant and amplitude measures indicate that speakers of Lopit may be provided with extra cues to the contrast. The role of such cues, in addition to duration, is worth consideration in typological work assessing the robustness of glide length contrasts in both perception and production.

6. Acknowledgements

I wish to thank members of the Lopit community for contributing their time and knowledge to this project, Janet Fletcher for advice and comments on this work, and Jonathan Moodie for illuminating discussions of Lopit morphosyntax. Support from ARC grant DP0984419, the University of Melbourne School of Languages and Linguistics, and the ARC Centre of Excellence for the Dynamics of Language is gratefully acknowledged.

7. References

[1] Maddieson, I., “Glides and gemination”, Lingua, 118:1926–1936, 2008.

[2] Kawahara, S., “Sonorancy and geminacy”, in L. Bateman, A. Werle, E. Reilly and M. O’Keefe [Eds], Univ. Massachusetts Occasional Papers in Linguistics 32: Papers in Optimality Theory III, 145-186, GLSA, 2007.

[3] Aoyama, K. and Reid, L. A., “Cross-linguistic tendencies and durational contrasts in geminate consonants: An examination of Guinaang Bontok geminates”, J. of the Int. Phonetic Assoc., 36(2):145–157, 2006.

[4] Billington, R. “‘Advanced Tongue Root in Lopit: Acoustic and ultrasound evidence”’, in J. Hay and E. Parnell [Eds], Proceedings of the 15th Australasian Int. Speech Science and Technology Conf., 119–122, ASSTA, 2014.

[5] Vossen, R., The Eastern Nilotes: Linguistic and historical reconstructions, Dietrich Reimer Verlag, 1982.

[6] Turner, D., Lopit phonology, SIL-Sudan, 2001.

[7] Stirtz, T., Phonological comparison of Lopit dialects, SIL-South Sudan, 2014.

[8] Ridouane, R., “Gemination at the junction of phonetics and phonology”, Laboratory Phonology 10:61–90, 2010.

[9] Hamzah, H., “The acoustics and perception of the word-initial singleton/geminate contrast in Kelantan Malay”, PhD thesis, Univ. Melbourne, 2013.

[10] Kawahara, S., “The phonetics of *sokuon*, or geminate obstruents”, in H. Kubozono [Ed], Handbook of Japanese phonetics and phonology, 43–78, Mouton, 2015.

[11] Ham, W. H., Phonetic and phonological aspects of geminate timing, Routledge, 2001.

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

[13] Cohn, A.C., Ham, W. H., and Podesva, R. J. “The phonetic realization of singleton-geminate contrasts in three languages of Indonesia”, in J. J. Ohala [Ed], Proceedings of the 14th Int. Congress of Phonetic Sciences, 587–590, Univ. California Press, 1999.

[14] Khattab, G. and Al-Tamimi, J., “Geminate timing in Lebanese Arabic: The relationship between phonetic timing and phonological structure”, Laboratory Phonology 5(2):231–269, 2014.

[15] Hansen, B.B., “The perceptibility of duration in the phonetics and phonology of contrastive consonant length”, PhD thesis, Univ. Texas at Austin, 2012.

[16] Boersma, P. and Weenink, D., Praat: Doing phonetics by computer [Computer program]. Version 5.2.35. Online: http://www.praat.org/, accessed 28 August 2011.

[17] Espy-Wilson, C.Y., “Acoustic measures for distinguishing the semivowels /w j r l/ in American English”, J. of the Acoustical Soc. of America 92(2):736–757, 1992.

[18] Harrington, J. & Cassidy, S., Techniques in Speech Acoustics, Springer Science & Business Media, 1999.

[19] Hunt, E. H., “Acoustic Characterization of the Glides /j/ and /w/ in American English”, PhD thesis, MIT, 2009.

[20] Cassidy, S. and Harrington, J., “Multi-level annotation in the Emu speech database management system”, Speech Commun. 33:61–77, 2001.

[21] R Core Team, R: A language and environment for statistical computing [Computer program]. Version 2.15.0 and 3.2.5. Online: http://www.R-project.org, accessed 16 April 2012/22 April 2016.

[22] Bates, D., Maechler, M., Bolker, B., and Walker, S., “Fitting Linear Mixed-Effects Models Using lme4”, J. of Statistical Software, 61(1):1–48, 2015.

[23] Moodie, J., “A grammar of Lopit”, PhD thesis, Univ. Melbourne, in progress.

[24] Billington, R. “Lexical tone in Lopit”, in The Scottish Consortium for ICPhS 2015 [Ed], Proceedings of the 18th Int. Congress of Phonetic Sciences, paper 0755. 1-4, Univ. Glasgow, 2015.

Child Kriol has stop distinctions based on VOT and Constriction Duration

Rikke L. Bundgaard-Nielsen[1], Brett J. Baker[2], Elise A. Bell[3]

[1]MARCS Institute for Brain, Behaviour and Development,
Western Sydney University, Australia
[2]University of Melbourne, Australia
[3]University of Arizona, USA

rikkelou@gmail.com, bjbaker@unimelb.edu.au, elisebell@email.arizona.edu

Abstract

We present acoustic analyses of stop productions by 16 Kriol-speaking children from two communities in the Northern Territory, Australia. Kriol has been characterised as having a variable phonological inventory and lexical items, presenting children with a difficult language-learning task. Our results suggest, on the contrary, that these children have canonical lexical specifications, and also indicate that their experience with L2 English in a school setting has not resulted in a shift towards more English-like Voice Onset Time and Constriction Duration settings. Indeed, the results are consistent with recent adult Kriol data and indicate that Kriol phonology is stable and shows no obvious evidence of decreolisation.

Index Terms: Kriol, VOT, constriction duration, first language acquisition.

1. Introduction

North Australian Kriol (hereafter Kriol) is an English-lexified creole spoken in Northern Australia by approximately 20,000 people [1], making it the most widely spoken Indigenous language (after English/Aboriginal English). Kriol is a recent contact language, having developed in the past 100 years [2] [3] [4] [5]. Many Indigenous communities are still undergoing language shift to Kriol [2] [4].

Kriol has often been described as exhibiting high degrees of phonological and lexical variation both within and between speakers [5] [6] [7]. Such a scenario would constitute an unusual and potentially very challenging 'moving target' for children acquiring L1 Kriol from birth: how do children identify word-learning targets and attune to their L1 phonological system in an environment of substantial (and often contradictory) phonetic and phonological variation?

This phonological and lexical variation has been described as falling along a 'creole continuum' [5] [6] [8] ranging from substrate-like to English-like, resulting in the absence of, for instance, stop voicing contrasts and fricatives at the basilectal end, and presence of those stop contrasts as well as (some) fricatives at the acrolectal (English) end of the continuum [6].

It has been argued that this variation is the result of a process of decreolisation as (increasing) interaction between creole speakers and speakers of the superstrate language leads to the introduction of lexical variants and constructions (more) consistent with the lexifier norm [5]. It is also argued that the relative position of an individual along the continuum at any given time is explained by reference to the speaker's conversational partner and sociolinguistic context [5].

Recently, [10] have argued that 'variability' in Kriol is not a result of decreolisation. Rather, they argue that observed 'variation' can be ascribed to three factors:

a) the unpredictable relationship between cognate Kriol and English lexemes, with respect to voicing and constriction;
b) differences in VOT boundaries between Kriol and English, leading to erroneous perception of [k] for some instances of Kriol /g/, for instance; and most importantly;
c) L2 speakers of Kriol participating in earlier research.

This view is consistent with recent experimental work with L1 adult Kriol speakers, suggesting in fact a stable phonological inventory identical to neither the substrate languages nor to English. According to [4] [10] [11], adult L1 Kriol, like English, has two series of stops, distinguished via long-lag/short-lag VOT [12] in initial position. Unlike English, however, Kriol word-medial stops are distinguished primarily by constriction duration (CD) differences of approximately 50-100ms (voiceless > voiced stops), depending on place of articulation [10]. This is similar to CD differences between 'fortis' and 'lenis' stops in some of the substrate languages: Ngalakgan, Ngandi, and Ritharrngu [13] [14]. English word-medial voiced and voiceless stops only differ minimally in terms of CD [15] [16], and this distinction is not perceptually relevant in connected speech [16]. See Figures 1 and 2 for (American) English and Kriol VOT and CD based on [15] and [10], respectively.

Given the role of English in the mainstream community and particularly in the schooling of Kriol-speaking children in the NT, the question of whether children's language is shifting towards the lexifier and away from the current adult L1 Kriol norm is of great theoretical and practical importance both within creolistics and from a second language/bilingualism perspective. Creolistics would frame a shift from L1 Kriol VOT/CD implementation towards more English-like VOT/CD implementation as decreolisation, while a second language/bilingualism perspective would frame such a shift (in one direction or the other) within an L1-to-L2 or L2-to-L1 transfer framework. Patterns of shift away from an L1 and towards an L2 phonetic stop voicing specification have been amply documented in the second language literature (see for instance [17] [18] [19] [20] [21]).

Similar patterns of drift or shift have also been demonstrated in studies of bilingual children, including Spanish-German bilinguals [22], and in very recent studies of Greek-German bilingual children, who, in both cases, must negotiate distinct phonetic realisations of stop voicing contrasts. Greek and Spanish have a distinction realised as pre-voiced vs short-lag aspirated stops, while the German distinction like English is short-lag aspirated vs long-lag aspirated [23]. Greek-German bilingual children enrolled in a German-language school produce more Greek voiced stops with German voicing setting, than comparable Greek-German bilingual children enrolled in Greek language school, and conversely those enrolled in Greek school similarly produce more Greek-like voiced stops in German [23]. Similar, but not identical, patterns have been observed for early Japanese-English bilingual children [24].

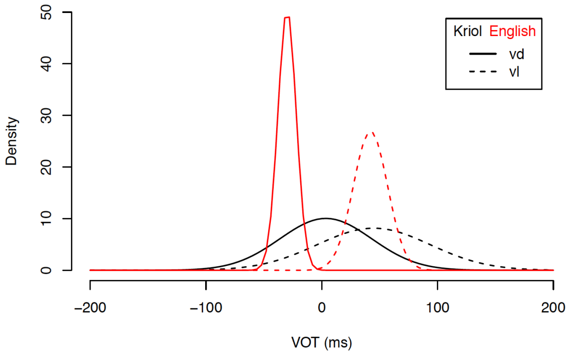


Figure 1. *Density curves of the distribution of VOT in voiced and voiceless stops in Kriol (black) and English (red). Fully drawn lines = voiced stops, dotted = voiceless stops. English and Kriol VOT based on [15] and [10], respectively.*

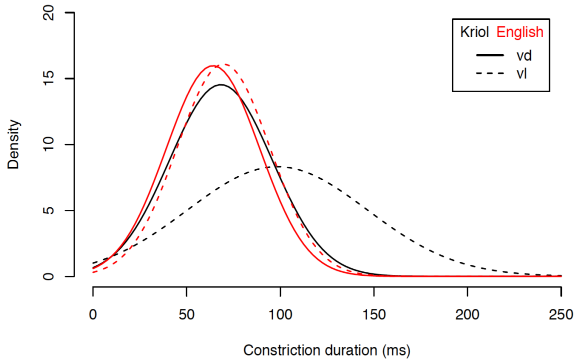


Figure 2. *Density curves of the distribution of CD in voiced and voiceless stops in Kriol (black) and English (red). Fully drawn lines = voiced stops, dotted lines = voiceless stops. English and Kriol CD based on [15] and [10], respectively.*

Finally, the present research addresses previous observations that child speech is more variable than adult speech, particularly in subsegmental elements such as VOT [25] [26] [27] [28]. Such variability does not, however, necessarily obscure language-specific consonant boundaries. Rather, [29] concludes that children produce the same contrasts as adults *'despite also exhibiting more variability in their production of individual vowels*' (p 1).

The question of excess variability of course remains pertinent with respect to consonants, and particularly interesting in the context of a language for which the claim has been made that even adult users are highly variable. In the following section, we describe our method for eliciting stop productions from children acquiring L1 Kriol.

2. Method

2.1. Participants

We recruited 16 children (9 female, range = 4;8 to 10;0) for participation in the present study. 13 participants lived in Beswick/Barunga and three in Numbulwar, in the Northern Territory, Australia. All were L1 Kriol speakers and attended the preschool program (1 participant) or the foundation (grades 0-2) program (15 participants) at Wugularr Primary School or Numbulwar School. All children were acquiring English as a second language (L2) at school: teaching at Wugularr and Numbulwar is conducted in English by mainly non-Indigenous teaching staff. Indigenous, Kriol-speaking support staff are, however, often present.

All children were reported to have normal hearing, but we did not conduct hearing screening in order to verify this. As recurrent otitis media is common in many Indigenous communities, it is possible that some of the children may have had some undiagnosed hearing impairment. The participants were recruited by word of mouth, through the first and second authors' existing contacts in the community. Parental consent and child assent was attained for all participants.

2.2. Materials

The recordings for the acoustic study of Kriol stop production by primary school-aged children were based on the selection of 24 easily depictable Kriol nouns, such as *door /duwa/*, *book /buk/*, *cat /ket/,* for which a highly salient and easily recognisable photograph was selected (See Figure 3).

/baik/	/kofi/	/bakit/	/buk/
/taiga/	/do:g/	/kerit/	/ket/
/babul/	/pigi/	/be:d/	/wota/
/te:dul/	/teibul/	/tiʃu/	/tut̪/
/spaida/	/geko/	/duwa/	/sebɹa/
/balu:n/	/pelikin/	/epul/	/guwana/

Figure 3. *Examples of pictures used in the elicitation task, and full list of elicited target words in IPA.*

To ensure that each of the 24 items would be known to children, the items were selected in consultation with two literate L1 Kriol speakers, one of whom is trained in early childhood education, and who has previously been involved with the creation of Kriol literacy materials for preschool and early primary school children. The 24 items all contained voiced and voiceless stops /p t k b d g/ in either initial or medial position, or both, in a wide range of vowel contexts. We avoided consonant clusters wherever possible.

2.3. Elicitation procedure

We elicited the Kriol lexical items in the following way: each participant was seated at a table in a quiet room in front of a PowerPoint presentation containing pseudo-randomised repetitions of the 24 pictures. Each picture was displayed with the pre-recorded Kriol prompt *Wanem dijan?* ('What is this?'), spoken by an unfamiliar native female Kriol speaker. The task was explained to the children in Kriol, as well as in English.

When a child provided a correct response, the experimenters gave positive feedback. Incorrect responses (for instance 'bible' for 'book' or 'water' for 'bottle') received feedback about the desired name. Each picture was displayed until the child responded, or until it was clear that no response would be given, at which point the item was skipped.

Responses were recorded using a PMD660 Marantz flash-RAM digital recorder with a DPA d:fine headset microphone. All recordings had a 16-bit sampling depth with a sampling rate of 44.1 kHz.

2.4. Acoustic analyses

We extracted a total of 1388 VOT measurements (950 word-initial and 438 word-medial) and 438 word-medial CD measurements. Some target items were produced multiple times in association with one presentation of a target; repetition did not disqualify tokens from analysis.

Tokens which were incomplete or interrupted by background noise, laughter, or contact with the microphone were excluded from analysis. Only items where the child produced the intended utterance (or a semantically related word which included the original target consonant in the intended position: 'cup' for 'coffee', 'pig pig' or 'bird bird' for 'pig' or 'bird', 'sea turtle' for 'turtle') were included in this analysis. Excluding non-target responses resulted in a loss of 3.1% of the VOT data and 1.98% of the CD data.

The acoustic recordings were analyzed in *Praat*. VOT was defined as the time between the burst/release of the relevant stop and the onset of voicing. Voicing measurements were taken at the zero crossing before the second clear periodic wave. Constriction duration was measured as beginning at the initial stop closure (the end of the preceding vowel's clear F2) and ending at the stop burst.

2.5. Predictions

The present study tests three competing hypotheses, H1, H2, and H3. According to **H1**, Kriol remains highly variable in terms of VOT contrast production, and Kriol children will produce highly variable VOTs and CDs between Kriol lexical items, as well as highly variable VOTs and CDs within Kriol lexical items (e.g. between repetitions of the same lexical item by a given child, and in the VOT and CD settings implemented for a given lexical item by different children). Such a pattern of behaviour would be attributable to an inherently unstable target language, because speakers move between basilectal and acrolectal inventories/lexical targets.

According to **H2**, child Kriol reflects current adult Kriol differentiation of stops, and child Kriol speakers will implement VOT and CD contrasts similar to VOT and CD contrasts in the adult Kriol input. Any variation observed should be commensurate only with the variability normally found in child language. In other words, child Kriol will provide evidence for a series of voiced and a series of voiceless stops, differentiated by VOT and CD.

Finally, under the assumption that Kriol is undergoing a process of decreolisation resulting in variation in VOT and CD, **H3** would predict that child Kriol would exhibit some variation in VOT and CD realisation, but any convergence in the VOT and CD measurements would be due to convergence with the norm for the acrolectal variety or the superstrate (English). In other words, the VOT and CD values will be somewhat variable but tending towards an English-like differentiation of voiced versus voiceless stops.

3. Results

3.1. Group results

The group results are presented in *Figure 4* (VOT) and *Figure 5* (CD) below. On the whole, the results are consistent with **H2**: A series of independent *t*-tests of voiced vs voiceless VOT were significant ($p < .001$) for all contrasts in word-initial position. Word-medially, *t*-tests of voiced vs voiceless VOT were significant ($p < .001$) for /k g/ and /p b/. For CD, *t*-tests were significant ($p < .001$) for /p b/, /t d/, and /k g/.

Together, these results indicate that young Kriol-speaking children use VOT and CD information to differentiate Kriol stops /p b/, /t d/, and /k g/ in a way that is very similar to that of adult Kriol speakers (Figures 1 and 2). There is no indication in this dataset that child Kriol speakers are acquiring a target language in which VOT and CD contrast maintenance is optional. Nor does the importance and consistent use of CD information indicate that these stop contrasts have been acquired through contact with L2 English: it is clear that Kriol children produce voiceless stops differing markedly from voiceless stops in English as VOT is the primary cue in English while Kriol relies also on CD (Fig. 2).

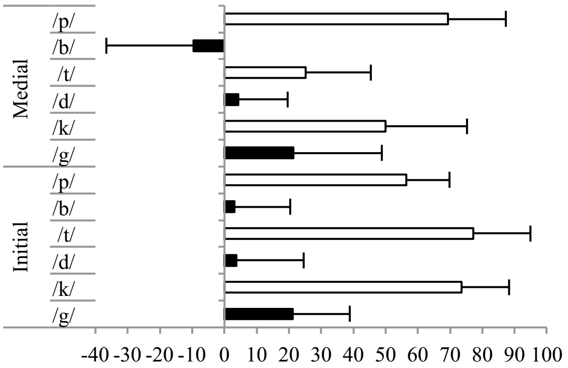


Figure 4. *Group means (ms) for word-initial and -medial VOT. Error bars indicate SD (plus values for positive entries, minus values for negative entries)*

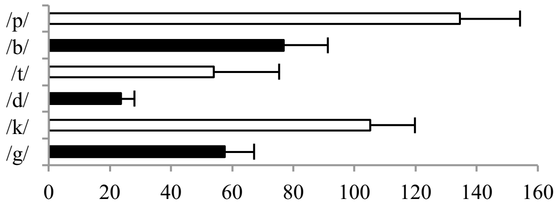


Figure 5. *Group means (ms) for word-medial CD measures. Error bars indicate SD (plus values for positive entries).*

3.2. Individual results

In order to examine whether the global results reported above were in fact characteristic of the Kriol stop production of all the children in this study, we conducted individual analyses for each child, collapsing across place of articulation. Word-initially, *all* 16 children produced voiced versus voiceless stops with systematically different VOTs (p = .003 or less in each case). Word-medially, four children (ages 5;4, 5;7, 6;3; 7;0) did not provide enough tokens for individual analysis. *T*-tests of the remaining 12 children's medial VOT and CD productions showed that five children maintained a contrast in both VOT and CD, while another five produced a VOT distinction, and four a CD distinction. Two children (B3 [5;4] and B12 [7;0]) did not produce a medial contrast in VOT or CD. As is evident from Figures 6 and 7, however, all children with *non-significant* individual results, however, produced voiceless VOT and CD values consistently longer than their voiced counterparts, except B2 (who produced a CD contrast).

Interestingly, the children who failed to produce a medial VOT distinction were not the younger cohort (5;2, 5;4, 6;1, 7;0 of age), so we find it unlikely that lack of a VOT-based distinction reflects a developmental trend, such that older children have acquired the word-medial distinction but younger children have not. Further, given the clear trend in the VOT and CD values among the children whose results were not significant, we find it plausible that the lack of significant individual differences in VOT and CD is due to lower numbers of medial tokens in the study. Finally, the fact that

CD appears to be a primary cue to the Kriol medial /p b/, /t d/, and /k g/ contrasts indicates that these contrasts are indeed Kriol and not transferred from L2 English into Kriol.

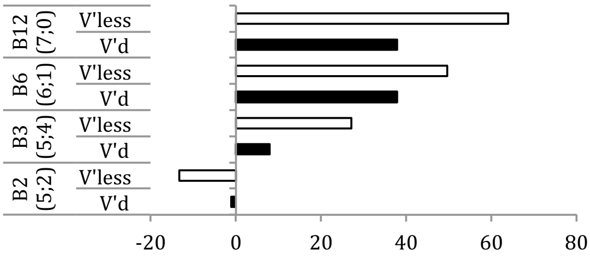


Figure 6. *Individual VOT means (ms) for children with non-significant t-test results. Child age in parenthesis. B indicates that the participants were from Beswick.*

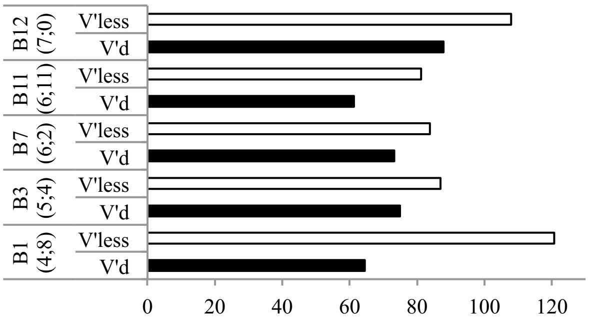


Figure 7. *Individual CD means (ms) for children with non-significant t-test results. Child age in parenthesis. B indicates that the partipants were from Beswick.*

4. General discussion

There is no doubt that the Kriol-speaking communities in Northern Australia, such as Beswick and Numbulwar, have been undergoing a complex process of language shift for a number of decades [2] [4]. This has contributed to pervasive views of Kriol as an inherently variable language, presenting unique language learning challenges for its users.

However, the effects of this pattern of language shift on the sound system of Kriol have not been instrumentally examined until now. This study demonstrates that the Kriol-speaking children tested here form a largely homogenous group, exhibiting little variation in stop voicing production. Moreover, the characteristics of their stops are not different from adult Kriol but differ from English stops. From these results we can infer the following. Kriol (in these communities) is not undergoing a rapid change in the production of obstruents sufficient to cause a difference between two generations of speakers. Kriol is also not approaching English; that is, is not decreolising, at least in the case of obstruent production. In addition, given that the adults recorded by [10] were from Ngukurr and Numbulwar, and the children discussed here from Beswick and Numbulwar, the results also strongly suggest that there is not a major dialect difference between these communities.

5. Acknowledgements

We thank Mandy Manggurra, Lenny Joshua, Hilda Ngalmi, the children who participated in this study, and Dr. Stephen and Joanne Hill. We are grateful for funding from the Australian Research Council (DP130102624).

6. References

[1] AIATSIS/Commonwealth of Australia, The National Indigenous Languages Survey Report, Dept of Comm, IT & Arts, 2005.

[2] Munro, "Roper River Aboriginal language features in Australian Kriol: Considering semantic features," in Lefebvre (Ed), Creoles, their Substrates, and Language Typology, 461–87, Benjamins. 2011.

[3] Harris, Northern Territory Pidgins and the Origin of Kriol, Pacific Ling., 1986.

[4] Bundgaard-Nielsen & Baker, "Fact or furphy? The continuum in Kriol," in Meakins & O'Shannessy (Eds) Loss and Renewal: Australian Languages Since Contact, 177-216, De Grutyer Mouton, 2016

[5] Sandefur, "An Australian Creole in the Northern Territory: A Description of Ngukurr-Bamyili Dialects," SIL, 1979.

[6] Sandefur, "Papers on Kriol: The Writing System and a Resource Guide," SIL, 1984.

[7] Schultze-Berndt et al., "Kriol," in Michaelis, Maurer, Haspelmath & Huber (Eds), The Survey of Pidgin and Creole Languages, vol. 1, 241–251, OUP, 2013.

[8] DeCamp, "Toward a generative analysis of a post-creole continuum," in D. Hymes (Ed), Pidginization and Creolization of Languages, 349-370, CUP, 1971.

[9] Rickford, Dimensions of a Creole Continuum, SUP, 1987.

[10] Baker et al., "The obstruent inventory of Roper Kriol," Aust. J. Ling. 34:307-344, 2014.

[11] Bundgaard-Nielsen & Baker, "Perception of voicing in the absence of native voicing experience," Proc. Interspeech, Dresden, 2015.

[12] Lisker & Abramson, "A cross-language study of voicing in initial stops: Acoustical measurements," Word 20:384-422, 1964.

[13] Baker, Word Structure in Ngalakgan, CSLI, 2008.

[14] Butcher, "'Fortis/lenis' revisited one more time: the aerodynamics of some oral stop contrasts in three continents," Clin. Ling. Phon. 18:547-557, 2004.

[15] Byrd, "54,000 American stops," UCLA Working Papers in Phonetics 83:97-116, 1993.

[16] Crystal & House, "Segmental durations in connected-speech signals: Syllabic stress," JASA 83:1574-1585, 1988.

[17] Flege, "Age of learning affects the authenticity of voice-onset time (VOT) in stop consonants produced in a second language," JASA 89:395-411, 1991.

[18] Flege & Frieda, "Amount of native language (L1) use affects the pronunciation of an L2," J.Phon. 25:169-186, 1997.

[19] Fowler et al., "Cross language phonetic influences on the speech of French–English bilinguals," J.Phon., 36:649-663, 2008.

[20] Antoniou et al., "Language context elicits native-like stop voicing in early bilinguals' productions in both L1 and L2," J.Phon., 38:640-653, 2010.

[21] Sancier & Fowler, "Gestural drift in a bilingual speaker of Brazilian Portuguese and English," J.Phon., 25:421-436, 1997.

[22] Kehoe et al., "Voice onset time in bilingual German-Spanish children," Bilingualism: Lang. & Cog. 7:71-88, 2004.

[23] Chionidou & Nicolaidis, "Voice onset time in bilingual Greek-German children," Proc. ICPhS, 2015

[24] Hareda (2003). L2 Influence on L1 Speech in the Production of VOT. Proc, ICPHS, Barcelona, 2003..

[25] Eguchi & Hirsh, "Development of speech sounds in children," Acta Oto-Laryngologica, Supp., 257, 1969.

[26] Tingley & Allen, "Development of speech timing control in children," Child Devel., 46(1), 1975.

[27] Whiteside et al., "Patterns of variability in voice onset time: a developmental study of motor speech skills in humans," Neurosci. Lett., 347:29-32, 2003.

[28] Yu et al., "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.

[29] Redford & Oh, "Fixed temporal patterns in children's speech despite variable vowel durations," Proc. ICPHS, 2015.

Lacking Vision: Insights into the Automatic Classification of Emotion in AMCs The Walking Dead

Joanne Quinn

Montclair State University
Montclair, NJ
quinnj11@montclair.edu

Abstract

Speech Emotion Recognition (SER) is in huge demand in our high-tech world, but can SER detect emotion with near-human accuracy? To explore this question, we must first explore others: What is near-human accuracy in SER? And: How much is that accuracy influenced by visual prosody? This study consists of two parts: The first contrasts the difference in emotional perception in near-natural speech when audio is presented alone, then in conjunction with visual stimuli. The results create a baseline for human auditory SER, which is used to judge a basic automatic SER classification model using prosodic, semantic, and temporal features.

Index Terms: speech emotion recognition, prosody, emotion detection

1. Introduction

Emotions in speech have been studied over decades and across disciplines. This research has provided a basic framework for selecting features that differentiate among strong emotions [1]. However, emotion is present in different levels of language structure, so a one-size-fits-all guideline for feature selection may be an overly simplistic model. Since emotion is implied, rather than entailed, and shaped by both culture and cognition, emotional encoding and decoding are complex processes that are also heavily speaker and listener dependent [2]. While verbal prosodic cues are believed to have patterns that can communicate emotion within and across languages and cultures, visual prosodic cues are also believed to convey information about the speaker's emotional state and temper [3],[4].

With advances in computational ability, researchers in many fields have turned their attention to automatic speech recognition (SER). Many of the features previously identified have proven useful in automatic SER, however, an optimal feature set for automatic recognition has not yet been established [5]. Additionally, the corpora used for SER are not necessarily representative of realistic emotions. Past research has included work with the Switchboard and Fisher corpora, prescribed, acted corpora, and fixed utterances in stories designed to elicit specific emotions [6],[3],[1]. Schuller et al. notes that "the types of emotions that normally are prompted are definitely not the same as one would encounter in realistic scenarios", and Rakov and Rosenberg criticize that the language used among strangers is much more formal than language used among friends [7],[6].

In order to address both the role of visual prosody in the interpretation of emotion and the limitations of past corpora, the first half of this study focuses on assembling and evaluating a corpus of acted, yet authentic speech. The second half of the study explores the effectiveness of common prosodic and semantic features in the automatic classification of SER with Weka's RandomForest algorithm.

2. Motivation

Graf et al. studied visual prosody in the facial movements of speakers and concluded that it is identifiable in the speech of most people, and, while varying from person to person, it is strongly correlated with the prosodic structure of the text. Although Graf et al. focused mainly on visual prosody of the face and head (the scope of their research was confined to modeling authentic talking heads), they do note that body language is used to facilitate turn taking and emphasize point of view [4]. For this research, the definition of visual prosody is expanded to include any gesture or movement that the speaker performs whilst he/she is speaking.

Rakov and Rosenberg explored the use of sarcasm with clips from MTV's animated series *Daria*. *Daria* was chosen for this task because it is not a traditionally acted corpus and it is rich in sarcasm. Furthermore, the authors felt that the animated nature of the show would lend to a more exaggerated expression of sarcasm [6]. In keeping with the idea that scripted television more closely mirrors human emotion, the corpus for this experiment was collected in the same manner.

The main prosodic feature selection was inspired by prior work in SER. Many studies have emphasized the importance of fundamental frequency in the detection of emotion. Additional features such as formant values and formant bandwidth values have been judged useful in automatic SER [8]. Additionally features related to energy and speech rate are calculated [9],[5].

Finally, the semantic features selected in this study were inspired by the norms of valance, arousal, and dominance collected by Warriner et al. [10].

3. Materials

3.1. Corpus creation

The corpus for this project is comprised entirely of clips from AMC's *The Walking Dead*. Evaluating realistic emotion is the goal of this study, and *The Walking Dead* was chosen for this task because, despite its fantastical premise, it has been nominated for numerous awards. In 2015, episodes of *The Walking Dead* occupied all 10 chart spots for the top 10 most watched scripted cable television shows. These facts imply that both critics and the general public alike believe the acting to be quite realistic, even if the situation is not as believable.

In total, 152 clips were selected from seasons 1, 2, 4, and 5 and recorded as avi files in November 2015. The clips were selected for the target emotions: “Happy/playful”, “neutral”, “sad/upset”, and “angry”. Context was considered in the judgments, and subtitles were recorded on screen. The final breakdown of stimuli per emotion for the researcher-selected clips was (N: 152):

- Happy/Playful: 45
- Neutral: 13
- Sad/Upset: 47
- Angry: 47

3.2. Corpus evaluation

In January and February of 2016, 4 male and 4 female audio/visual raters were recruited to view and rate all 152 clips. The raters ranged in age from 20 to 37.

The clips were randomized and a Java program was written that allowed a rater to replay a clip until he/she decided on an emotional category. Participants were not provided with a definition or an example of the emotions.

After all evaluations were completed, the inter-rater reliability for each stimulus was calculated using Fleiss’ Kappa. Any item that did not receive a kappa score of 0.31 or above was removed from consideration. Additionally, any item that was evenly judged between 2 or more categories was removed. The final breakdown of stimuli per emotion after the audio/visual rating phase was (N:134):

- Happy/Playful: 35
- Neutral: 34
- Sad/Upset: 38
- Angry: 27

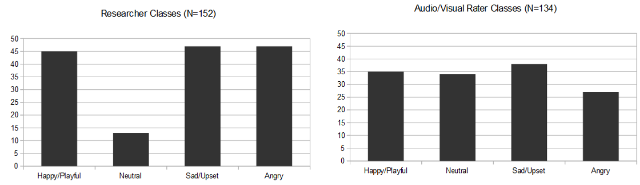


Figure 1: Researcher Selected Classes vs AV Rater Labeled Classes

4. Methods

4.1. Experiment 1

Between January and April, 2016, Amazon’s Mechanical Turk, an online marketplace for human intelligence tasks, was used to survey the 28 audio-only participants. MP3 and ogg (audio) files were created for each of the 152 stimuli and text files were created with the transcribed audio of each clip. Participants were asked to listen to the audio while reading the text and then assign each clip to an emotional category. Again, participants were not provided with a definition or an example of the emotions. The survey took approximately 50 minutes to complete.

4.1.1. Evaluation

Although the audio-only raters were surveyed on all 152 stimuli from the original set, their results were pre-processed to remove the ambiguous clips and those that did not receive a kappa score of .031 or above in section 3.2.

Responses were judged on a binary scale, receiving a score of 1 if the correct category was identified and a score of 0 if an incorrect category was identified.

4.1.2. Results

Overall, the 28 audio-only respondents correctly identified the stimuli 59.68% of the time. Previous research by Banse and Scherer has reported accuracy rates of around 55% on similar tasks; additionally the authors propose a recognition rate of around 50% as a stable estimate of acoustic emotional recognition rate [11]. Below is the combined confusion matrix for all 28 audio-only respondents. Note that, while the majority class of the stimuli was “sad/upset”, the majority of respondents selected “neutral” for every class, excepting that of “happy/playful”.

a	b	c	d	← Classified As	
5	10	16	4	a	= Happy/Playful
3	13	11	7	b	= Neutral
7	14	6	11	c	= Sad/Upset
7	9	6	5	d	= Angry

Table 1: Confusion Matrix for Audio Participants

The percent of correctly identified stimuli in this experiment was used as a benchmark to evaluate an automatic SER classifier.

4.2. Experiment 2

4.2.1. Feature selection

All stimuli in the corpus were manually annotated in Praat [12]. Values for the word and sentence level features described below were then automatically extracted using a series of Praat scripts. In the pre-processing step, audio was converted to mono, and each file was denoised with Praat’s denoising feature. In total, each stimulus was represented by a 50 item feature vector, though some features were used for normalization, not classification. At this point, 5 additional stimuli were removed from the corpus because automatic extraction of features was unsuccessful. The final corpus contained 129 items labeled as such:

Category	**Count**	**Percent**
Happy/Playful	32	24.8
Neutral	33	25.6
Sad/Upset	37	28.7
Angry	27	20.9
Total	129	100

Table 2: Final Class Totals

The majority class was “sad/upset”, comprising 28.7% of the data. No effort was made to ensure that speaker genders were evenly distributed. At this point, the data was split by speaker gender (two clips spoken by young children were categorized as “female”):

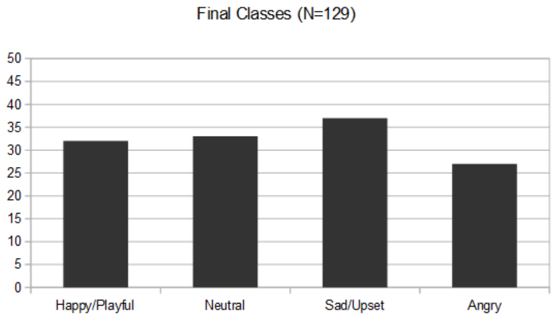


Figure 2: Final Classes for Automatic Evaluation

Gender	Count	Percent
Female	60	46.5
Male	69	53.5

Table 3: Speaker Gender

4.2.2. Word-level and temporal features

A Praat script was created to iterate through each word in each file, extracting the average measurement for that word. Word level features include:

- Average Word Length
- Speech Rate
- F0: Average, Minimum, Maximum
- Intensity: Average, Minimum, Maximum
- Changes in pitch and intensity were calculated from the ranges

4.2.3. Sentence-level features

Another Praat script was written to iterate through all files, extracting average measurements for each utterance. For all averages, log(10) was also calculated. Utterance features included:

- Average formant values and formant bandwidths for the first, second, and third formants
- Word Count
- F0: Average, Minimum, Maximum
- Intensity: Average, Minimum, Maximum
- Total Change in Pitch and Intensity
- Absolute value of change in Pitch and Intensity
- Average F0 and Intensity were separately normalized over the average for all utterances and the gender-average for all utterances

4.2.4. Semantic features

Bash scripting was used to assign overall affect scores to each utterance using the theory of arousal, valance, and dominance. Each word of each utterance was first lemmatized using the Stanford Core NLP Toolkit [13]. The Lemmas were then compared to a database of word norms collected by Warriner et al. [10]. Total raw scores for valence (the pleasantness of a stimulus), arousal (the intensity of emotion provoked by a stimulus), and dominance (the degree of control exerted by a stimulus) were calculated. Each of these values was then calculated as a percent over the total affect value of the utterance.

- Arousal/TotalAffect
- Valance/TotalAffect
- Dominance/TotalAffect

5. Results and discussion

For the classification task, in order to test the features and select the best subset of them, Weka's CorrelationAttributeEval was run on the full set of features [14]. Using the output of the algorithm, a total of 18 features were chosen for classification.

Selected Features
Speech Rate
log(10) F1, F2, and F3 Bandwidth
log(10) F2 and F3
Average Pitch Utterance over Average for all samples
Average Pitch Utterance over Average for gender
Log(10) Average Pitch
Max pitch over average
Change in Pitch and Intensity
Min and Max intensities over average intensity
Each affect value as a percent over all

Table 4: Final Selected Features

Weka's RandomForest classifier was most successful in classifying emotions into their labeled categories [15]. This result was surprising as none of the prior research consulted in this study made use of decision trees for classification. This classifier was run via the Weka GUI using 10-fold cross validation. The algorithm creates a collection of decision trees, which then vote for the most popular class. Overall, the RandomForest classifier attained an accuracy of 40.31%. Precision, recall, and F-measures for each class are reported in table 5 below.

Precision	Recall	F-Measure	Class
0.267	0.250	0.258	Happy/Playful
0.414	0.364	0.387	Neutral
0.417	0.541	0.471	Sad/Upset
0.545	0.444	0.490	Angry

Table 5: Precision, Recall, and F-Measures by Class

The confusion matrix in table 6 shows the algorithm's correctly and incorrectly classified instances for each emotion. A comparison of this matrix to the one in Table 1 above, reveals that the algorithm correctly classified every class except "neutral" more accurately than did the human, audio-only respondents. (The audio-only respondents correctly identified "neutral" utterances 38.2% of the time, whereas the algorithm achieved a correct classification 36.4% of the time; however, the audio-only respondents also incorrectly classified most "sad/upset" and "angry" utterances as "neutral".) One possible explanation for the audio-only respondents majority "neutral" classification is that they may have defaulted to this class when they were unsure of which other class to select.

A problem for both the automatic SER and the audio-only respondents was correctly classifying the "happy/playful" emotions. Both experiments have a large number of "happy/playful"

a	b	c	d		← Classified As
8	6	13	5	a	= Happy/Playful
10	12	9	2	b	= Neutral
7	7	20	3	c	= Sad/Upset
5	4	6	12	d	= Angry

Table 6: Confusion Matrix for Classification

items classified as “sad/upset”. There are three possible reasons for this:

The first reason is the actual construction of the classes. While minimization of classes was intentional in this research, it may have benefited the outcome to have more fine-grained class definitions. For example, the class of “sad/upset” comprises both the low arousal, sad, and the high arousal, upset. Splitting this class into two separate classes and providing a high arousal and low arousal option for the “happy/playful” class (e.g. “contentment” and “excitement”) may solve some of the classification errors. It is possible that people heard excitement or laughter and confused the emotion with being upset or crying.

Another possible reason for these classification errors is that, in the hopes of experimenting on near-natural speech, exaggerated speech emotion was avoided, making the classification task harder than it would be on a traditionally acted corpus.

Finally, visual prosody also may have played a role in the incorrect classification of the “happy/playful” stimuli, especially in regard to the playful stimuli. Sometimes, when one person is joking with another person, that playfulness may be conveyed with a smile or a wink. While the audio-visual participants were able to evaluate the utterance in context with the visual prosody, these visual cues were unavailable to the audio-only respondents who may have mistaken the playfulness for another emotion.

While the overall automatic classification rate is still rather low and did not attain the benchmark goal of 59.68%, it does represent an increase of 11.6% over a majority-class baseline, while also correctly identifying 3 of the 4 classes more frequently than the audio-only respondents did. Additionally, it is important to consider that the rate was attained without the use of short-term spectral features such as linear prediction cepstrum coefficients (LPCC) and mel-frequency cepstrum coefficients (MFCC), which have both shown success in speech recognition algorithms.

6. Conclusion

This paper aims to advance the field of automatic SER by first providing a baseline of how well humans can identify emotion without visual prosody in near-human, acted speech. The creation and categorization of the new *The Walking Dead* Emotion Corpus was crucial in the task of evaluating human competency in SER. The automatic classification of that corpus based only on prosodic and semantic features not only provides a glimpse into which features are important for automatic SER, but it also highlights an important classification method, decision trees, which may deserve more consideration in future work. Improvements to this corpus could be made with the addition and evaluation of more data and the partitioning of classes into high arousal/low arousal variants of each emotion. Improvements on the automatic SER classification rates may be achieved with the addition of the short-term spectral features (such as MFCC), which were discussed earlier.

7. Acknowledgments

I would like to thank Dr. Jonathan Howell, Dr. Anna Feldman, and Dr. Jin Peng of Montclair State University for their helpful guidance and suggestions, all the participants who donated their time to make this study possible, and the reviewers who offered thoughtful, thought-provoking feedback.

8. References

[1] Sobin, C. and Alpert, M., “Emotion in speech: The acoustic attributes of fear, anger, sadness, and joy.”, Journal of Psycholinguistic Research, 28: 347-65, 1999.

[2] Majid, A., “Current Emotion Research in the Language Sciences.”, Emotion Review, 4: 432-443., 2012.

[3] Baum, K.M. and Nowicki, S.N., “ Perception of emotion: measuring decoding accuracy of adult prosodic cues varying in intensity.”, Journal of Nonverbal Behavior, 22: 89-106., 1998.

[4] Graf, H.P., Cosatto, E., Strom, V., and Huang, F.J., “Visual prosody: Facial movements accompanying speech.”, Paper presented at the 5th International Conference on Automatic Face and Gesture Recognition, Washington, DC., 2012.

[5] Vogt, T., Andre, E., and Bee, N., “Emovoice - a framework for online recognition of emotions from voice.”, OC. of an IEEE Tutorial and Research Workshop on Perception and Interactive Technologies for Speech-Based Systems., 5078, 188-99., 2008.

[6] Rakov, R. and Rosenberg, A., “Sure, I Did The Right Thing: A System for Sarcasm Detection in Speech.”, Proceedings of the 14th Annual Conference of the International Speech Communication Association, Lyon, France, 2529 August 2013.

[7] Schuller, B., Batliner, A., Steidl, S., and Seppi, D., “Recognising realistic emotions and affect in speech: state of the art and lessons learnt from the first challenge.”, Speech Communication, 53(9/10), 10621087., 2010.

[8] Petrushin, V.A., “Motional Recognition in Speech Signal: Experimental Study, Development, and Application.”, ICSLP-2000, 2: 222-5., 2000.

[9] Joshi, A., and Kaur, R., “A Study of speech emotion recognition methods.”, Int. J. Comput. Sci. Mob. Comput.(IJCSMC), 4: 28-31., 2013.

[10] Warriner, A.B., Kuperman, V. and Brysbaert, M., “Norms of valence, arousal, and dominance for 13,915 english lemmas.”, Behavior Research Methods, 44(4)., 2013.

[11] Banse R. and Scherer, K., “Acoustic profiles in vocal emotion expression.”, Personality Social Psych, 70(3): 614-636., 1996.

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

[13] Manning, C.D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S.J., and McClosky D., “The Stanford CoreNLP Natural Language Processing Toolkit.”, Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations, 55-60. 2014.

[14] Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I.H., “The WEKA Data Mining Software: An Update.”, SIGKDD Explorations, 11(1)., 2009.

[15] Breiman, Leo., “Random Forests.”, Machine Learning, 45(1):5-32. 2001.

Depression Prediction Via Acoustic Analysis of Formulaic Word Fillers

Brian Stasak[1,2], *Julien Epps*[1,2] *and Nicholas Cummins*[3]

[1] School of Elec. Eng. & Telecomm., University of New South Wales, Sydney, Australia
[2] Data61-CSIRO, Sydney, Australia
[3] Chair of Complex and Intelligent Systems, University of Passau, Passau, Germany

brian.stasak@student.unsw.edu.au, j.epps@unsw.edu.au, nicholas.cummins@uni-passau.de

Abstract

Understanding what kind of speech is most effective for predicting depression deserves more attention since diagnosis and monitoring is often based on limited interview or questionnaire responses. Consequently, this paper investigates thin slice data selection using token word analysis, demonstrating that spoken formulaic language and in particular filler words hold acoustic discriminative properties that are useful when predicting different ranges of depression. Results from an analysis of the DAIC corpus indicate that filler words are equally or more effective for depression prediction as using entire utterances, and that acoustic and linguistic features combined generate competitive depression prediction results.

Index Terms: depression, formulaic language, speech emotion, thin slice.

1. Introduction

Continuous digital bio-signal analysis is widely used in many fields of healthcare, such as neurology, cardiology, and physiology. In recent years, speech has gained attention as a bio-signal measure aiding mental health diagnoses and monitoring. Automated speech processing, although not fully implemented as a standard protocol tool in assessing and monitoring depressed individuals, is not far off in the future as a potential clinical application. The measurement and analysis of vocal characteristics is advantageous over other bio-signals because of its naturalistic communicative form and non-invasive collection without complex expensive machines that require specialized training (e.g. fMRI, PET, SQUID).

Notwithstanding the ease of collecting speech recordings, the dilemma of knowing which vocal segments carry compact informational value is a questioned predicament by researchers. The concept of 'thin slice' data selection was studied in [1], wherein brief social or clinical observations can often yield useful compact information rather than longer observations. Researchers surmise several considerations regarding thin slice data selection theory: (1) the channel of observations, whether verbal or non-verbal, has little effect on the predictive observational results; (2) a great deal of behavioral affect is generated unintentionally or unconsciously yet it still contributes to other people's observed predictions or interpretations; (3) when affective thin slice data selection is proven effective it can significantly reduce resources without sacrificing performance; and (4) while examining human behavior the thin slice approach works particularly well when predicting vital interpersonally oriented criterion variables [1].

The concept of data selection has proven applicable in many areas of speech processing. Prior research indicates in speech processing applications, such as speech recognition, speaker identification, and speech emotion classification, data selection that reduces phonetic variability and compares similar phonetic structures can improve performance. For instance, Reynolds et al. [2] demonstrated by focusing on specific acoustic classes (i.e. text-dependent words) individual speaker models develop better modeling of short-term variations thereby resulting in higher overall identification performance for short utterances. In Boakye et al. [3], researchers analyzed thirteen token words consisting of less than ten-percent of all utterances across speakers. Their results showed speaker identification performance was superior for short fixed token words than for an entire set of words.

Researchers have also advocated for the inclusion of linguistic features with acoustic features in paralinguistic applications. For instance, Shriberg and Stolcke [4] and Ishihara et al. [5] analyzed text transcripts to generate token word linguistic features that contributed to higher speaker recognition performance. Research has also shown habitual speaker idiolects and formulaic language use can provide unique speaker information [3]. Formulaic language is common in everyday discourse and even monolog narratives, where it contributes to nearly a quarter of all conversational speech [6]. It includes conventional word expressions, proverbs, idioms, expletives, hedges, bundles, and fillers.

In brief, depressed speaker characteristics reported to date include: reduction in vocal prosody dynamics; decreased vocal pitch; reduced vocal intensity; slower rate of speech; increased vocal tenseness; motor incoordination, including disfluency or motor retardation [7]; reduced vowel space ranges [8]; and difficulty with word retrieval [9]. Further, investigation into spoken phonetic variability across individual phonemes has indicated potential biomarkers for depression disorders which are based on speech rate and phoneme durations [10].

In the literature, only a limited number of *non-acoustic* linguistic text-based studies have specifically examined formulaic word fillers in depressed populations [6, 11]. The *acoustic* evaluation of formulaic language for predicting levels of depression has several practical advantages. Its related filler words are found spontaneously occurring in large numbers across a wide range of speakers irrespective of gender, age, language, and education. Furthermore, the constrained acoustic phonetic variability that arises naturally from considering only token filler words facilitates intra- and inter-speaker comparison. *Intra-speaker*, there are typically multiple examples of each token word per utterance, helping to construct a more focused analysis of phoneme characteristics. *Inter-speaker*, token word characteristics can also be more readily compared between speakers than entire utterances. Studies have demonstrated that even a single phoneme type in large quantities can help to reveal discriminative information regarding a speaker's emotional state [8, 12].

While formulaic language has arguably only a minor contribution to semantic/pragmatic content, it adds importantly to cognitive-emotive speaker internalization [6]. In this paper, it is hypothesized that since formulaic fillers are quite common in number and represent speakers' mental internalization, they will reveal more about the effects of depression on speech than utterances in general.

2. Speech Corpus

The audio portion of the training and development from the Distress Analysis Interview Corpus (DAIC) [9][1] was used for all experiments herein. The DAIC was designed to investigate language, nonverbal behaviors, psychophysiology, and assisted human-computer commutative dialog. This database was chosen because it provides a fixed set of utterances which were spoken by a computer generated interviewer; a large group of speakers, 143 males and females; high-quality close-talking microphone recordings; natural speech in a clinical type environment; and PHQ-8 evaluations along with scores per participant. The PHQ-8 is a popular eight-questioned self-administered mental health assessment tool commonly used in diagnosis of depression disorders [13]. It has an interval scale of 0 to 24, where larger scores imply a greater depression severity. The DAIC was also chosen because it includes phrase-level transcripts with beginning/ending time markers, which made extracting single token words possible with minimal error. Only individually segmented token word entries were evaluated. These segments were determined based on a transcriber's indication of when single word tokens began and ended. For more information regarding the DAIC transcription conventions see [9]. The token words evaluated in all experiments are listed in Table 1. The ten token words combined included 95% of speakers from the train and 100% of speakers from the development sets; the other 5% were omitted due to transcript time marker errors.

Table 1: *Description of token words evaluated, percentage of training/test speaker coverage, total number of utterances, and number of unique speakers in the DAIC.*

Token Words	Word Type	% Train	% Test	# Total	# of Unique Speakers
"Hmm"	Filler	33%	49%	139	52
"Mhm"	Filler	35%	43%	123	53
"Mm"	Filler	48%	60%	168	72
"Uh"	Filler	52%	60%	298	77
"Umm"	Filler	85%	94%	1276	124
"So"	Filler	27%	54%	119	48
"No"	Polar	77%	74%	230	109
"Yeah"	Polar	60%	66%	305	88
"Okay"	Polar	27%	43%	56	44
"You Know"	Bundle	21%	23%	57	31

3. Experimental Methodology

3.1. Feature Overview

The baseline experiments used the 88 eGeMAPS acoustic features [14]. Additionally, there were 116 acoustic features extracted using VoiceSauce [15]. Examples of the eGeMAPS and VoiceSauce functional features include median and standard deviations for jitter, shimmer, Mel-Frequency Cepstral Coefficients (MFCC), pitch, formant frequencies, formant bandwidths, formant amplitudes, and harmonic-to-noise ratios. For all acoustic features, windows of 20 ms (with 10 ms overlap) were applied.

Linguistic features were derived from individual speaker's entire recording session transcript and compiled using text-processing scripts. Although many linguistic features were considered, the average utterance length, average syllables per second, percentage of unique words, percentage of articles/prepositions/pronouns, and readability scores per speaker were most valuable. Readability scores were based on common methods, such as Flesh-Kincaid Grade Level and Gunning Fog Index [16]. While these readability scores are typically derived from written passages, they can also be useful when qualitatively applied to verbal transcripts.

3.2. System Design

The AVEC 2016 depression prediction sub-challenge baseline [17] utilised all training and development data in the DAIC, applied similar acoustic features, and employed a support vector machine for regression analysis. The baseline acoustic features were created using entire utterances, and a depression prediction baseline of 5.35 Mean Absolute Error (MAE) and 6.74 Root-Mean Squared Error (RMSE) was achieved. For comparison with this baseline, Support Vector Regression (SVR) was also used to predict the depression scores for experiments herein. SVR has been successfully applied to speech depression/emotion prediction tasks and is known for effective statistical generalization [18]. Based on the SVR output, two standard performance metrics were used to evaluate the overall predictive accuracy due to their application in recent speech depression prediction challenges: MAE and RMSE. One distinct advantage the RMSE has over the MAE is it does not utilize absolute values and is generally better at revealing model performance differences.

In Figure 1, the system design involves two main inputs, speech and spoken transcript data. During the acoustic and linguistic feature extraction stage, feature selection can be performed (indicated by shaded boxes). The feature selection process retains the most salient features and omits any weaker features for statistical predictive analysis. Note that occasionally some speakers' token words were too short in verbal duration, so some acoustic features could not be adequately calculated (producing nulls) and/or some linguistic features were found to have little variance. Afterwards, a depression score prediction output was generated and compared to the ground truth depression PHQ-8 score.

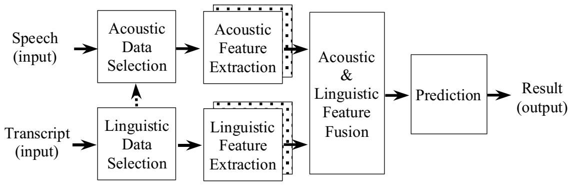

Speech (input)
Acoustic Data Selection
Acoustic Feature Extraction
Transcript (input)
Linguistic Data Selection
Linguistic Feature Extraction
Acoustic & Linguistic Feature Fusion
Prediction
Result (output)

Figure 1: *Acoustic data selection involves chosen token words, whereas linguistic data selection uses all words. Feature selection was applied during the feature extraction stages.*

[1] Quality control listening per utterance was completed to check time-stamp accuracy (some transcriptions had time-stamp errors). File labels and their time-stamps are available per request via the authors.

4. Results

4.1. Token Words Versus Entire Utterances

The purpose of these experiments was to examine how smaller segments perform when compared with entire utterances, and determine which features or feature combinations contribute to better token word depression prediction. Formulaic filler word analysis nearly matched that of the entire utterances baseline when identically partitioned comparisons were made, as shown in Table 2. When evaluated against the entire utterances baseline, fillers gave lower overall MAE and RMSE.

Table 2: *Token word and entire utterances (baseline) depression prediction using eGeMAPS acoustic features and SVR.*

	Word/Phrase		Baseline (all utt.)	
Token Words	MAE	RMSE	MAE	RMSE
"*Hmm*"	3.85	5.07	5.22	6.90
"*Mhm*"	4.08	4.18	4.10	4.80
"*Mm*"	5.58	6.95	5.70	6.77
"*Uh*"	5.37	6.50	5.96	6.90
"*Umm*"	6.56	8.15	5.35	6.67
"*So*"	6.31	8.07	6.31	8.17
"*No*"	5.00	6.08	4.71	5.79
"*Yeah*"	7.17	8.87	6.10	7.10
"*Okay*"	5.73	6.74	5.55	6.54
"*You Know*"	8.67	9.92	3.07	4.31
All Average	**5.83**	**7.05**	**5.21**	**6.40**
Filler Average	**5.29**	**6.49**	**5.44**	**6.70**

Similar results were found for VoiceSauce features, but with lower error in general. These features may have performed better due to containing a greater number of total features than eGeMAPS. In addition, VoiceSauce applies more than one acoustic analysis method (i.e. Praat, Straight, Snack Sound Toolkit) for estimating frequency and energy formant-related features. VoiceSauce features, shown in Table 3, produced competitive results when compared with the entire utterances baseline.

Table 3: *Token word SVR depression prediction using VoiceSauce acoustic features*.

	Word/Phrase	
Token Words	MAE	RMSE
"*Hmm*"	2.91	4.13
"*Mhm*"	3.76	4.62
"*Mm*"	5.56	7.09
"*Uh*"	4.82	6.03
"*Umm*"	6.27	7.96
"*So*"	6.06	8.57
"*No*"	4.87	6.24
"*Yeah*"	7.38	10.91
"*Okay*"	5.08	6.68
"*You Know*"	7.52	8.96
All Average	**5.42**	**7.11**
Filler Average	**4.90**	**6.40**

Using these features, only one of the token words fillers "*umm*" achieved a higher MAE than that of the baseline. Note that "*hmm*" performed particularly well for depression prediction, generating a 2.91 MAE versus the 5.22 MAE in the baseline.

4.2. Linguistic Baseline System

Trends in the linguistic features were discovered for depressed speakers having higher range PHQ-8 scores (e.g. 15-23). For instance, depressed speakers tended to have a reduction in overall word syllable averages, reduced preposition usage, increased usage of pronouns, and overall simpler sentence structure based on average readability scores. While depressed versus healthy female speakers did not indicate a difference in average words per sentence, depressed males showed an overall reduction, especially for higher PHQ-8 scores. In experimenting with linguistic features the MAE and RMSE average using linguistic features was nearly equal to the entire utterances baseline acoustic functional features, 5.17 and 6.30, respectively.

4.3. Acoustic and Linguistic Features Combined

Experiments utilizing all acoustic features from token words along with linguistic features were completed in an attempt to attain the lowest the MAE and RMSE possible. The eGeMAPS, VoiceSauce, and linguistic features were concatenated as a single vector per utterance before prediction using SVR. In Table 4, experiments using a combination of acoustic token word and linguistic features produced the overall lowest MAE and RMSE average for fillers token words.

Table 4: *Token words SVR depression prediction using combined eGeMAPS, VoiceSauce, and linguistic features*.

	Combined	
Token Words	MAE	RMSE
"*Hmm*"	2.89	4.35
"*Mhm*"	3.45	4.63
"*Mm*"	5.90	7.13
"*Uh*"	5.00	6.32
"*Umm*"	6.18	7.82
"*So*"	5.05	7.09
"*No*"	5.06	6.31
"*Yeah*"	6.42	8.50
"*Okay*"	4.70	6.24
"*You Know*"	8.08	9.42
All Average	**5.27**	**6.78**
Filler Average	**4.75**	**6.22**

In the results presented to this point, only subsets of the training/test data could be used. To understand the depression prediction performance across the entire data, all token words were merged into training and test sets, which allowed for every speaker to be represented much like the baseline results found in [17]. Using the combined entire utterances baseline eGeMAPS features, prediction errors of 5.51 MAE and 6.83 RMSE were attained. When combined sets were then run using the filler words with eGeMAPS features, prediction errors of 6.07 MAE and 7.52 RMSE were achieved.

While the combined filler word results did not produce results as low as the entire utterances baseline, filler words are surprisingly accurate considering most comprise less than a second of speech. The combined filler word results may be an indication that some particular filler words and their acoustic-phonetic attributes are better for depression prediction than others.

4.4. *N*-Best Analysis

An *n*-best approach was experimented with, using the four lowest MAE/RMSE token words that, when combined, allowed score prediction for every test utterance; thus, creating a fair comparison with the entire utterances baseline. In Table 5, the best test MAE/RMSE performance was achieved using *n*-best eGeMAPS and linguistic features with feature reduction. The absolute improvement in MAE/RMSE over the entire utterances baseline was 0.95 and 1.24, respectively. The *4*-best token words ("hmm", "mhm", "no", "uh") were fillers and/or had nasal phonetic elements. Due to experimental time constraints, entire utterances baseline for VoiceSauce MAE/RMSE will be included in a later revised version. For token words, the VoiceSauce features attained similar results to eGeMAPS. However, they did not demonstrate further improvement with the addition of linguistic features and feature reduction.

Table 5: *Comparison of entire utterances baseline versus 4-best token words ("hmm", "mhm", "no", "uh") on all test speakers. Note * indicates feature selection applied.*

	eGeMAPS		VoiceSauce	
	MAE	RMSE	MAE	RMSE
All Utterances (similar to [17])	5.51	6.83	-	-
All Fillers	6.07	7.52	6.08	7.59
4-Best	4.72	5.76	4.71	5.71
4-Best + Linguistic	4.74	5.70	4.71	5.71
4-Best* + Linguistic*	4.56	5.59	4.71	5.71

5. Conclusion

This research demonstrates that thin slice speech data selection can be competitive for depression prediction when compared with using whole utterances. Moreover, results show that among the token words selected for study herein, fillers consistently provided the lowest depression score prediction error.

Filler words appear advantageous because they are naturally repeated in abundance. The general location of filler words is between phrase clauses; meaning they typically begin or end a phrase, making them potentially easier to identify with automatic speech recognition and/or keyword spotting applications. Future research, utilizing a larger set of filler words with equal counts and number of speakers could further help determine which specific fillers or phonetic content contains the most valuable speech information for depression prediction systems.

6. Acknowledgements

The work of Brian Stasak and Julien Epps was partly funded by ARC Discovery Project DP130101094. The work of Nicholas Cummins is supported by the EC's 7th Framework Programme through the ERC Starting Grant No. 338164 (iHEARu).

7. References

[1] Ambady, N., & Rosenthal, R., "Thin slices of expressive behavior as predictors of interpersonal consequences: a meta-analysis", *Psych. Bulletin*, Vol. 111, No. 2, 256-274, 1992.

[2] Reynolds, D., & Rose, R., "Robust text-independent speaker identification using gaussian mixture speaker models", *IEEE Transactions on Speech and Audio Processing*, Vol. 3, No. 1, 72-83, January 1995.

[3] Boakye, K., & Peskin, B., "Text-constrained speaker recognition on a text-independent task", *ODYSSEY '04*, The Speaker and Language Recognition Workshop, 2004.

[4] Shriberg, E., & Stolcke, A., "The case for automatic higher-level features in forensic speaker recognition", *INTERSPEECH*, Brisbane – Australia, 1509-1512, 2008.

[5] Ishihara, S., & Kinoshita, Y., "Filler words as a speaker classification feature", *SST 2010*, Melbourne – Australia, 34-37, 2010.

[6] Bridges, K., "Prosody and formulaic language in treatment-resistant depression: effects of deep brain stimulation", PhD Thesis, Steinhardt School of Culture, Education, and Human Development: NYU – USA, 2014.

[7] Cummins, N., Scherer, S., Krajewski, J., Schnieder, S., Epps, J., & Quatieri, T., "A review of depression and suicide risk detection and assessment using speech analysis", *Speech Communication*, Vol. 71, 10-49, 2015.

[8] Scherer, S., Lucas, G., Gratch, J., Rizzo, A., and Morency, J., "Self-reported symptoms of depression and PTSD are associated with reduced vowel space in screening interviews", *IEEE Trans. Affect. Comp.*, Vol. 7, 59-73, 2016.

[9] Gratch, J., Artstein, R., Lucas, G., Stratou, G., Scherer, S., Nazarian, A., Wood, R., Boberg, J., DeVault, D., Marsella, S., Traum, D., Rizzo, S., & Morency, L., "The distress analysis interview corpus of human and computer interviews", *LREC*, 3123-3128, 2014.

[10] Trevino, A., Quatieri, T., & Malyska, N., "Phonologically-based biomarkers for major depressive disorder", *EURASIP Journal on Advances in Signal Processing*, Vol. 42, 2011.

[11] Pope, B., Blass, T., Siegman, A., & Raher, J., "Anxiety and depression in speech", *Journal of Consulting and Clinical Psychology*, Vol. 35, 128-133, 1970.

[12] Sethu, V., Ambikairajah, E., & Epps, J., "Phonetic and speaker variations in automatic emotion classification", *INTERSPEECH*, ICSA, Brisbane – Australia, 617-620, 2008.

[13] Kroenke, K., Strine, T., Spitzer, R., Williams, J., Berry, J., & Mokdad, A., "The PHQ-8 as a measure of current depression in general population", *Journal of Affective Disorders*, Vol. 114, 163-173, 2009.

[14] Eyben, F., Scherer, K., Schuller, B., Sundberg, J., André, E., Busso, C., Devillers, L., Epps, J., Laukka, P., Narayanan, S., and Truong, K., "The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing", *IEEE Transactions on Affective Computing*, in press, 2015.

[15] Shue, Y., Keating, P., Vicenik, C., & Yu, K., "VoiceSauce: a program for voice analysis", Proceedings of the *ICPhS* XVII, 1846-1849, 2011.

[16] Wu, D., Hanauer, D., Mei, Q., Clark, P., An, L., Lei, J., Proulx, J., Zeng-Treitler, Q., & Zheng, K., "Applying multiple methods to assess the readability of large corpus medical documents", *Student Health Technology Information*, 192, 647-651, 2013.

[17] Valstar, M., Gratch, J., Schuller, B., Ringeval, F., Lalanne, D., Torres Torres, M., Scherer, S., Stratou, G., Cowie, R., & Pantic, M., "AVEC 2016 – depression, mood, and emotion recognition workshop and challenge", submitted June 2016.

[18] Cummins, N., Sethu, V., Epps, J., & Krajewski, J., "Relevance vector machine for depression prediction", in Proceedings of the *Annual Conference of the International Speech Communication Association*, *INTERSPEECH*, Dresdan – Germany, 110-114, 2015.

Time to Embrace Emotion Change: Selecting Emotionally Salient Segments for Speech-based Emotion Prediction

Zhaocheng Huang[1,2] *and Julien Epps*[1,2]
[1]School of Electrical Engineering and Telecommunications, UNSW Australia
[2] Data61, CSIRO, Australia
zhaocheng.huang@unsw.edu.au, j.epps@unsw.edu.au

Abstract

Continuous prediction of emotion dimensions has gained popularity recently, because systems of this kind can capture subtle changes in emotions in naturalistic settings. However, most of these systems take utterance-level or frame-level features as input, without considering within-utterance variation in emotion. This paper investigates data selection for speech-based emotion prediction from an emotion change perspective, and finds that annotation delays vary even within utterances. Experimental results on the RECOLA corpus show that emotion-change frames carry relatively greater emotion-related information, achieving 5.4% and 24.5% relative improvements over the baseline for arousal and valence prediction under the Output-Associative Relevance Vector Machine Framework.

Index Terms: data selection, annotation delays, emotion changes, continuous emotion prediction, relevance vector machine

1. Introduction

Within the affective computing community, there has been a trend towards representing emotions in terms of arousal and valence dimensions, which are numerical values representing how activated a person is and how pleasant they feel [1]. They are considered a more descriptive representation of complex and subtle emotions in naturalistic environment, compared with conventional emotion categories, such as neutral and anger [2]. This trend has been further driven by the annual Audio/Visual Emotion Challenges (AVEC) [3], which targets continuous prediction of arousal and valence.

Continuous emotion prediction is a regression problem, which involves feature extraction and regression modelling. More specifically, features are extracted from training utterances and then used to train a regression model, based on which features extracted from testing data can be used to generate predictions. Among possible regression models, the Support Vector Regression (SVR) and Relevance Vector Machine (RVM) have been shown to be successful for this task [4]. In this work, RVM is preferred because an advanced framework based on RVM, called Output-Associative (OA) RVM, has shown promise on AVEC data [5]. The OA-RVM comprises two stages of RVM regression modeling. The first RVM is trained on input features. Temporal arousal and valence predictions from the first RVM are then associated with input features for RVM training at the second stage.

During regression modelling, it is well-recognized that there exist annotation delays in emotion ratings, which are manually assigned by annotators. The delay is mainly caused by reaction lag between annotators' perceptual observations and decision-making [6], as well as fatigue or variations in attention. This delay has a great impact on emotion prediction system performance, and correct annotation delay compensation has been associated with quite dramatic improvements in accuracy [4].

Although there have been extensive investigations into emotion recognition during last decade, most studies neglect within-utterance variation and treat all parts equally. This may not be a good assumption in general, and improvement in performances of speech-based emotion prediction systems requires a better understanding of emotionally-salient segments within speech. Attempts to investigate this previously include examining emotion recognition accuracies using specific phonemes or phoneme classes, where it was found in [7], [8] that vowels, especially /a/ are more conducive to emotion classification, whilst Bitouk et al. [9] suggested that spectral features extracted from consonants are more effective. Le et al. [10] examined various data selection strategies based on classifier agreement, and compared utterance selection with sub-utterance selection (segments within utterances) in terms of emotion classification performances, as well as convergence rate and stability in training process.

Kim [11] speculated that another way to identify emotionally salient segments is to explore variations in emotions: low variation in emotions implies clear expression of emotions, which may favor emotion recognition. However, this have not been experimentally investigated, and recent studies suggest that short-term emotion dynamics can facilitate emotion classification [12] and diagnosis of psychological diseases [13]. In continuously annotated emotional corpora, it is observed that emotion ratings tend not to change across time, which leads to a large proportion of frames without changes in emotion ratings. Motivated by [12], [13], as well as this latter observation, it is reasonable to pose the question: are frames with emotion change more salient for predicting emotions?

2. Database

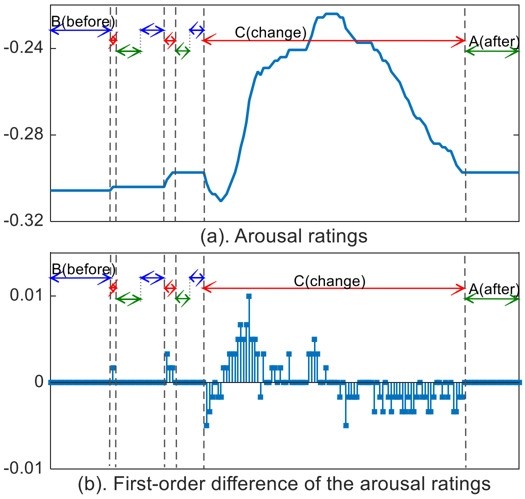
The database used in this paper is Remote Collaborative and Affective Interactions (RECOLA), a spontaneous multimodal corpus collected in settings where two French speakers remotely collaborate to complete a survival task via a video conference. During the collaborative interactions, multimodal signals, including audio, video and physiological signals such as ECG and EDA, were collected from 46 participants (data from 23 participants are publically available). This database is chosen because it is a large, high quality database that has been continuously annotated at every 40 milliseconds for arousal and valence by six annotators. Moreover, the recent

AV+EC2015 challenge [3] employed a subset of the database (18 speakers in total and 5-minute recording per speaker), which was evenly partitioned into training and development sets for a continuous emotion prediction task. In this study, we considered only speech signals and the same partitions as used in AV+EC 2015.

3. Data Selection

3.1. Defining Partitions based Emotion Change

This section defines different database partitions based on emotion changes, i.e. changes in the arousal and valence ground truth provided in the RECOLA corpus. To separate the “change” frames from “non-change” frames, first-order differences from the emotion ratings were calculated, as seen in Fig.1, where all data are partitioned into three parts: B (“before”), C (“change”), and A (“after”). This data partition scheme is applied to arousal and valence separately.



(a). Arousal ratings

(b). First-order difference of the arousal ratings

Figure 1: *All data are divided into B, C, and A based on the first-order difference where zeros mean no emotion change, whilst non-zeros mean emotion change. B and A were separated in order to understand differences between before and after emotion changes. These is a large proportion of “non-change” frames.*

As shown in Fig.1, partition C contains all the frames where emotion ratings change in addition to all frames whose ratings remain unchanged for less than *L* frames. Partition B contains all the frames before emotion ratings change at the beginning of every file, and the second half of all frames whose ratings remain unchanged for more than *L* frames. Similarly, partition A contains all the frames after the last change frame of every file, and the first half of all frames whose ratings remain unchanged for more than *L* frames. *L* is the minimum number of frames considered for non-change frames, i.e. B and A. This parameter is introduced to provide more continuity for C, as seen in Figure 1. With *L* = 1 frame, partitions B, C and A account for 16.86%, 63.18%, and 19.96% respectively for arousal. For valence, partitions B, C, and A account for 19.36%, 58.10%, and 22.54% respectively.

3.2. System Overview

In addition to feature extraction and regression modelling, the proposed system includes data selection and delay compensation. Data selection partitions all training data into three subsets: B, C and A. Delays are compensated via fixed temporal shifts and smoothing at the training and testing phase respectively, as per [4]. The regression model used in this paper is RVM, chosen because it offers good performance with fast training time. More importantly, it has shown promise across various system settings within the OA-RVM framework.

For training the RVM, we employed the *SparseBayes MATLAB toolbox*, and the only parameter to be tuned is the iteration number, selected from between 10 and 30. The temporal window size for construction of output-associative matrices containing input features and spatial temporal predictions was fixed to 151 frames as in [4]. The features used were 88-dimensional EGEMAPS functionals [3], extracted using a 2 second window size every 40 milliseconds, to align with the emotion ratings. All features were scaled into [0, 1] before training, and scaling coefficients from training data were used to normalize testing data. To ensure comparability, data selection was only conducted on training data, and all results are reported using all test data. During delay compensation, optimum delay values were chosen from [0, 6] s with 0.4 s increments. Performances of emotion prediction systems were measured using Concordance Correlation Coefficients (CCC) [3].

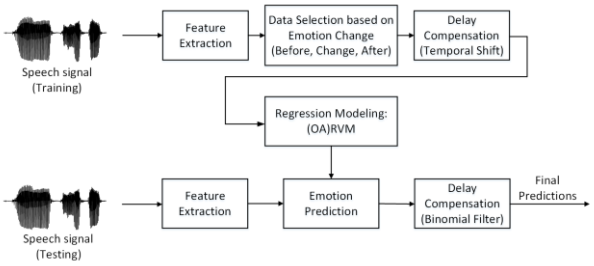

Speech signal
(Training)
Feature Extraction
Data Selection based on Emotion Change (Before, Change, After)
Delay Compensation (Temporal Shift)
Regression Modeling: (OA)RVM
Speech signal
(Testing)
Feature Extraction
Emotion Prediction
Delay Compensation (Binomial Filter)
Final Predictions

Figure 2: *Proposed system for investigating data selection based on emotion change.*

3.3. Emotion Prediction using Subsets of Training data

This section compares performances of different systems trained using either all training data or different subsets from Section 3.1, namely B, C, or A, all with a global delay compensated, tested on all test data (with no partitioning). The global delays, estimated from different delay values trialled on all training data and test data, as seen in Figure 2, were found to be 3.2s for arousal and 3.6s for valence.

Table 1: *Comparisons of performances using either all data (B+C+A) or subsets of training data (B, C, or A).*

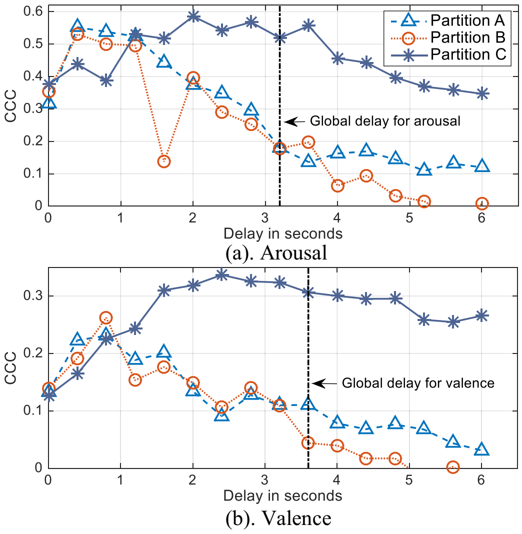
	B+C+A	B	C	A
Arousal	0.60	0.19	0.52	0.20
Valence	0.33	0.09	0.31	0.11

As shown in Table 1, the system trained on only change frames (C only) performed comparably to the system using all data (B+C+A), especially for valence. With the caveat that partition C contains a relatively larger amount of data than B or A, this suggests that frames where emotion ratings change carry more emotion-related information that favors emotion prediction.

Since annotation delay is important, and may vary between different partitions, it is perhaps unwise to keep a global delay value for different training partitions. This motivates us to search for the optimum delays for different training partitions (i.e. B, C and A).

3.4. Annotation Delay Optimized for Data selection

This experiment investigates how different delay values impact system performances when a subset of training data is used for training. The global delays estimated in section 3.2 were fixed for all test data (which remained unchanged throughout), whereas optimum delays for selected training partitions were trialled from among [0, 6] s with 0.4 s increments arousal and valence respectively. This approach was adopted throughout the following experiments.



(a). Arousal

(b). Valence

Figure 3: *Delay compensation for different data partitions (A, B and C), selected based on the first-order difference of arousal and valence ratings.*

As seen in Figure 3, the optimum annotation delays (for this database) were found to be around 0.4-0.8s for partitions B or A and around 2-3s for C, which were consistent for arousal and valence. This suggests that annotation delays are different not only among various annotators [6], but also different across time within-annotator. The greater delays for C are because of annotators taking more time to react to emotion change. Moreover, optimum delays for partition C are close to the global delays, suggesting that the delays mainly come from people's reactions to emotion change (C), not B or A, which is expected but, to the best of our knowledge, has not been shown in literature before. It also suggests that the global delays are influenced more strongly by change speech segments than non-change speech segments.

Systems trained on C only with optimum delay provided equivalent performances to systems trained on all data in Section 3.3. We repeated the above experiments using the OA-RVM framework, comparing systems trained on B+C+A with those trained on C only: 0.71 vs 0.71 for arousal, 0.40 vs 0.41 for valence. Since the OA-RVM framework consistently provided better performances than RVM (used in Fig. 3), the OA-RVM was used throughout all experiments below.

4. Emotion Change based Data Selection

4.1. Smoothed deltas vs First-order difference

In the above systems, we partitioned the training data based on the first-order differences of emotion ratings. However, the first-order difference is problematic because: (i) it assumes the possibility of extremely rapid emotion changes (i.e. the sampling interval between ratings is 0.04 seconds), which are unrealistic; (ii) raters tend not to move their cursor continuously, which results in a very large proportion of zero values, as shown in Fig.1(b); (iii) there is annotation noise caused by annotator tremble [14]. To resolve this, we proposed smoothing the first-order differences using a Moving Average (MA) filter, followed by applying a threshold to select "large emotion changes". The window size of the MA filter, herein referred to as W (measured in frames), needs to be chosen: the larger W, the smoother the changes in emotion ratings. In order to find the best W, we compared different values from 10 to 240 frames for arousal and valence respectively, as seen below in Fig.4.

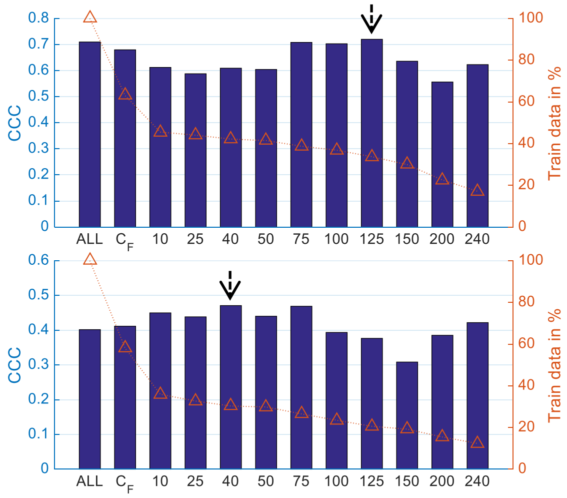


Figure 4: *"ALL" means baseline system trained on B+C+A, whereas "C_F" means a baseline system trained on only C (selected from the first-order difference). Other systems used C only based on smoothed deltas with different W values. The arrows indicate W values chosen to arousal and valence.*

A benefit offered by smoothed differences is that the change value for each single frame is a collective decision from all the frames within the smoothing window rather than only a change between two adjacent frames. This can reduce annotation noise, and a large value potentially suggests that the majority of frames within the temporal window are changing. To this end, smoothed differences are more beneficial than first-order differences for selecting emotion change. As shown in Figure 4, compared with first-order differences (C_F), smoothed differences provide better performances. The best W values for arousal and valence were 125 frames (CCC=0.72) and 40 frames (CCC=0.47) respectively, which were retained through the following experiments. This may suggest that arousal prediction is favored within a large region where emotion ratings in the majority of frames are changing, whereas valence prediction is more effective within smaller regions (10 - 75 frames).

Moreover, notice that although C_F, change frames based on first order difference, accounts for 63.18% and 55.10% of all the training data for arousal and valence respectively, systems with around 30-40% of the training data provided the best performances. This may suggest that large changes are more informative for emotion prediction, which motivates investigation into large emotion changes.

4.2. Large Emotion Changes

This section investigates emotion prediction systems trained on large emotion change frames. To do this, we applied a thresholds T to select large changes, which however leads to a reduction in C-partition training data. To mitigate this problem, we included adjacent frames around the large emotion changes for training. This leads to different combinations of *Ts* and number of adjacent frames around the large emotion changes, as seen in Figure 5.

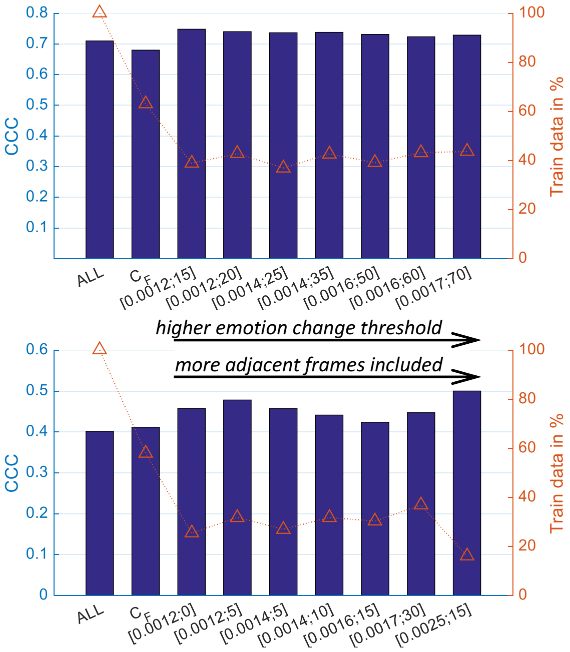


Figure 5: *Performances of systems trained on large emotion changes, compared with two baselines "ALL" and "C_F". [0.0012, 5] means selecting all emotion changes larger than T = 0.0012 and considering the nearest 5 frames adjacent to large changes. Percentages of resultant partitions with only large changes were shown in the orange markers.*

In Figure 5, we gradually increased thresholds to eliminate frames with small variations (which leads to less data) and included adjacent frames to maintain sufficient data for training (around 20% - 50% of total training data). It can be seen from Fig. 5 that systems trained on regions where large emotion changes occur yielded better performances over the baseline trained on all training data, achieving 0.75 vs 0.71 for arousal and 0.50 vs 0.40 for valence. Notice that this is achieved using 38.95% and 16.06% of training data for arousal and valence, lending support to the hypothesis that large emotion changes are more salient for emotion prediction.

Furthermore, in order to check how well the results generalise, the best settings for arousal and valence were tested via 6-fold cross-validation (15 speakers for training and 3 speakers for testing per fold). For arousal, W=125 frames, with [0.0012, 15] for large changes and 2.8 s delay for selected training data at each fold. For valence, W=40 frames, with [0.0012, 5] and 2.4 s delay for selected training data. System performances under these settings were 0.72 for arousal, and 0.41 for valence. The performances do not generalise very well. This is presumably due to the fixed delays, which are better to be optimized in training data within each fold.

5. Conclusions and Future Work

This paper has investigated data selection based on emotion changes for speech-based emotion prediction systems. Experimental results consistently show that speech segments containing emotion changes are more salient for emotion prediction (especially for valence). Training on only Change (C) frames gives comparable performances for arousal prediction and slightly better performances for valence, compared with performances using all data. When large emotion change (C) frames were used for training, after smoothing first-order differences, we achieved 5.4% and 24.5% improvements in CCC for arousal and valence relative to the baseline. This is significant because valence prediction from speech is generally recognized to be a difficult problem in the literature [15].

Moreover, this paper experimentally demonstrates that delays in emotion perception, reflected in annotation, mainly arise from people's reactions to emotion change (C), not non-change segments (B or A).

This paper is limited in that only one database has been tested. Future work involves extending this investigation, i.e. data selection based on emotion change, to multiple databases.

6. Acknowledgement

This work is partly funded by Data61, CSIRO.

7. References

[1] Gunes, H. and B. Schuller, "Categorical and dimensional affect analysis in continuous input: Current trends and future directions," *Image Vis. Comput.*, vol. 31, no. 2, pp. 120–136, 2013.

[2] Cowie, R., G. McKeown, et al., "Tracing Emotion," *Int. J. Synth. Emot.*, vol. 3, no. 1, pp. 1–17, 2012.

[3] Ringeval, F., B. Schuller, et al., "AV+EC 2015 – The First Affect Recognition Challenge Bridging Across Audio, Video, and Physiological Data," in 5th *AV+EC, ACM MultiMedia*, 2015.

[4] Huang, Z., T. Dang, et al., "An Investigation of Annotation Delay Compensation and Output-Associative Fusion for Multimodal Continuous Emotion Prediction," in *AVEC*'15.

[5] Nicolaou, M. a., H. Gunes, et al., "Output-associative RVM regression for dimensional and continuous emotion prediction," *Image Vis. Comput.*, vol. 30, no. 3, pp. 186–196, 2012.

[6] Mariooryad, S. and C. Busso, "Correcting time-continuous emotional labels by modeling the reaction lag of evaluators," *IEEE Trans. Affect. Comput.*, 2014.

[7] Lee, C., S. Yildirim, et al., "Emotion recognition based on phoneme classes.," in *INTERSPEECH*, 2004, pp. 889–892.

[8] Sethu, V., E. Ambikairajah, et al., "Phonetic and speaker variations in automatic emotion classification," in *INTERSPEECH*, 2008, pp. 617–620.

[9] Bitouk, D., R. Verma, et al., "Class-level spectral features for emotion recognition," *Speech Commun.*, vol. 52, pp. 613–625, 2010.

[10] Le, D. and E. Provost, "Data selection for acoustic emotion recognition: Analyzing and comparing utterance and sub-utterance selection strategies," in *ACII*, 2015.

[11] Kim, Y., "Exploring sources of variation in human behavioral data: Towards automatic audio-visual emotion recognition," *Affect. Comput. Intell. Interact. (ACII),* 2015.

[12] Provost, E., "Identifying salient sub-utterance emotion dynamics using flexible units and estimates of affective flow," in *ICASSP*, 2013.

[13] Houben, M., W. Van Den Noortgate, et al., "The Relation Between Short-Term Emotion Dynamics and Psychological Well-Being: A Meta-Analysis.," *Psychol. Bull.*, vol. 141, no. 4, pp. 901–930, 2015.

[14] Metallinou, A. and S. Narayanan, "Annotation and processing of continuous emotional attributes: Challenges and opportunities," *in 10th International Conference on FG* 2013.

[15] Grimm, M., K. Kroschel, et al., "Primitives-based evaluation and estimation of emotions in speech," *Speech Commun.*, vol. 49, pp. 787–800, 2007

Continuous Spoken Emotion Recognition Based on Time-Frequency Features of the Glottal Pulse Signal within Stressed Vowels

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

In speech production, emotional cues can be detected via three main aspects: excitation source, vocal tract and prosodic pattern. This paper addressed the first one, extracting six time and frequency related features from glottal pulse signals, transformed from stressed vowels. Four sustained vowels incorporating five regular emotions, which were selected from sentence recordings of the Berlin emotional speech database were investigated. The effectiveness of those glottal pulse features towards emotion recognition was proven through double round Robin quadratic classification in both single-gender and cross-gender stages, reaching average overall hit-rate of 63%, 64% and 53% for male, female and cross-gender respectively.

Key words: IAIF, glottal pulse, open quotient, speed quotient, frequency tilt, double round Robin Classification

1. Introduction

Most current human-machine speech communication systems are implemented with a simple one-channel interaction which merely transmits explicit verbal messages but lacks the capability of digging out the hidden intent, motive, and physiology state clues which speech may convey in the emotion of the speakers. Spoken information exchange cannot be fully achieved with only semantics. Some subtle acoustic characteristics embedded in emotion should also be captured to express intentions during speech synthesis.

According to the source-filter speech production model [1], speech can be viewed as the convolution of the glottal source, vocal tract filter, nasal cavity, lip radiation and articulation noise. Emotional clues have been extensively observed in various spectral features derived from the vocal tract (see [2] for an excellent summary of the studies). Several studies have also investigated some glottal-based features that are capable of classifying speech stress and emotion-related health confusions [3-5]. However, there is still not enough evidence compared with the vocal tract to address the contribution of the glottal source related features in differentiating emotion states. This is the main focus of this paper.

Iterative adaptive inverse filtering, IAIF [6] was applied in this study to derive glottal pulse signals from an emotional database. IAIF is a two-stage iteration process based on the principle of discrete all pole modelling (DAP) which recursively estimates the vocal tract model for every Hanning windowed 25ms analysis frame across every vowel. The glottal pulse signal is obtained by inverse filtering the vocal tract and lip radiation models from the original speech.

In this study six glottal pulse descriptive features are investigated in Section 2. Three are time-domain features, and they are mean open quotient (mOQ), standard deviation open quotient (stdOQ) and mean speed quotient (mSQ). Three are frequency-domain features, which are all methods of representing the spectral tilt. In Section 3 Principle Component analysis is used to reduce these six features to four dimension variables, which are combined as input vectors for subsequent classification model construction. Double round Robin classification [7] rather than the conventional support vector machines [8] was adopted in classification stages to examine the recognition performance, these results are discussed in Section 4. Conclusions are given in Section 5.

2. Extraction of glottal pulse features

2.1. Stressed vowel selection

The speech corpus used in this study is the Berlin emotional speech database Emo-DB [9] collected from ten (five male and five female) professional actors. Utterances were randomly sorted and based on ten linguistically neutral German sentences. Recording were stored with a 16 kHz sampling frequency as 16 bit numbers. There were seven emotions in the corpus: anger, boredom, disgust, fear, happiness, neutral and sadness, and there was an unequal distribution of utterances for each emotion. In this corpus the emotion of disgust achieved the worst subjective listening recognition rate (79.6%) [9] and the level of fear significantly differed from speaker to speaker. Therefore, for reliability and consistency of the recognition result, only exemplars of the remaining-five emotions were investigated in this study, yielding 725 sentences. Those emotional sentences were automatically labelled at word and phonetic level by the Munich Automatic Web Segmentation System, webMAUS [10]. All labelled sentences were converted into an EMU formatted database [11] and each vowel's onset and offset were extracted. To minimize potential errors in the detection of vowel boundaries, the first and last 5% of those extracted vowels would be removed before further processing.

In order to successfully recognize emotions from the glottal signal, we found that the length of selected vowel tokens must be over 65ms. In continuous speech only long tense monophthongs, such as **a**: can fulfil this constraint. Only four common sustained vowels /**a**: **o**: **i**: **e**:/ were examined in this study. It is natural that not all the extracted vowel tokens contain sufficiently distinguishable emotional clues. It is suggested that only those stressed vowels are able to clearly identify the emotional states of their corresponded sentence. To automatically identify stressed vowels three prosodic characteristics were examined for each vowel: (a) long duration, (b) changing pitch, (c) strong intensity. Any vowel token possessing at least two of above characteristics was deemed to be stressed. Thresholds were used to determine whether the three prosodic characteristics were present, however stressing thresholds altered with different emotions. For sadness and boredom the three thresholds were 80ms, 10Hz and 70rms respectively. For happiness and neutral the thresholds were 80ms, 10Hz, 75rms while for anger the values were 90ms, 15Hz, 80rms. The reason why higher threshold values were used for anger was to lower the number of selected vowel tokens. If this was not applied there would be a huge bias towards anger in the analysis dataset. Table 1 summarize the total number of available vowels and final selected stressed ones for each emotion and vowel type.

Table 1. *Available emotional vowels and identified stressed vowels (in parentheses) in Emo-DB*

	Angry	Sad	Neutral	Happy	Bored
e:	154(43)	68(43)	85(44)	81(48)	93(47)
a:	104(40)	60(40)	59(41)	54(39)	59(35)
i:	149(37)	68(40)	97(45)	74(33)	89(50)
o:	53(18)	29(22)	32(22)	27(18)	34(20)
sum:	460(138)	225(145)	273(152)	236(138)	275(152)

2.2. Time domain features of the Glottal Pulse

To properly parameterize the shape of glottal pulse waveform, three types of feature points must be accurately identified.

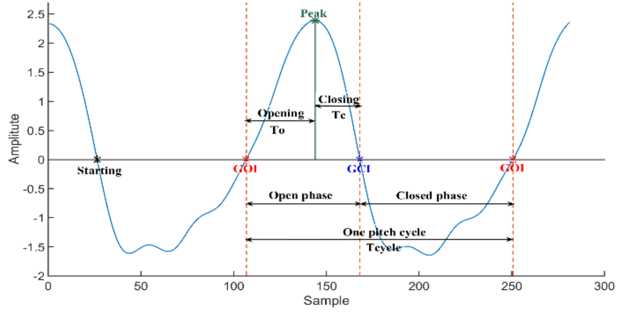


Figure 1: *Glottal pulse wave phases*

These are the glottal opening instance (GOI), the glottal closing instance (GCI) (both with an amplitude of 0), and the peak point between each GOI and GCI, indicating the maximal glottal opening when the glottal air flow at a maximum (see in Figure 1). The duration of two neighboring GCIs or GOIs represents one pitch cycle and the duration between each GOI and GCI pair can be regarded as the glottal open phase. In this study IAIF method has shown to give robust results of locating GOI and GCI independent of speakers, genders and emotions. Due to the sampling quantization limit that each two consecutive sample points have constant time spacing 0.0625ms, GCI and GOI points rarely coincide with sampled points, therefore linear interpolation between each two zero-crossing adjacent points is used for more precise GCI and GOI detection [12].

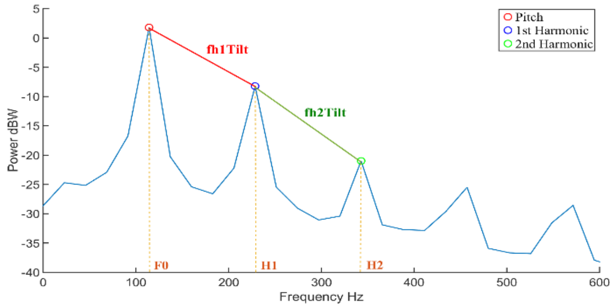


Figure 2: *Glottal pulse spectral tilt*

In order to derive the time domain features from the glottal signal it is important to identify the starting point first, which is the zero crossing prior to the first GOI (see Figure 1). The starting point is the very first detected GCI point rather than GOI since GCI theoretically correspond to the linear prediction residual peaks which are easier to find [13].The open quotient is the time ratio of when the glottal folds are open to the corresponding span of the pitch cycle and the speed quotient is given as the ratio of the opening phase over the closing phase.

$$OQ = \frac{T_o + T_c}{T_{cycle}} \tag{2}$$

$$SQ = \frac{T_o}{T_c} \tag{3}$$

To derive the glottal pulse and calculate the open quotient and speed quotient from each vowel segment, a sliding rectangle window of fixed length with a 1ms shift was used. The OQ is calculated in a portion of the waveform which is deemed to be stable. The analysis window moves until stability is reached, where the mean OQ values from the present and previous two windowed segments vary within a reasonably small margin. The window length and variance margin are adjusted according to input vowel length (see in Table 2), these values were determined heuristically. Typically there are between 4 to 16 glottal cycles in an analysis window. Results from the final window interval are collected to calculate the wanted mOQ, stdOQ, mSQ values.

Table 2. *Preset values for different length vowels*

Vowel length(ms)	Window length(ms)	Variance margin
(65,100]	45	0.02
(100,150]	60	0.01
(150,250]	80	0.006

2.3. Frequency domain feature

When converting signal from time domain to frequency domain, there is always a tradeoff between frequency resolution and noise-reduction [14]. In this study the main concern is noise, which has been introduced by both irregular glottal closures and extremely short sample lengths. Various frequency tilt values in the region between 0 to 3700 Hz have been investigated and successfully used in depression disorder identification [8]. Inspired by this, two different spectral tilts will be examined for emotion discrimination. Figure 2 gives an example of Welch transformed frequency plot of a glottal pulse signal converted from a /**a**: / token in the database. The spectral peaks can be easily identified and two lines can be well fitted to the first three peaks. Three frequency features depending on these peaks are defined as:

$$fh1Tilt = \frac{P(f_0) - P(h_1)}{f_0 - h_1} \tag{4}$$

$$fh2Tilt = \frac{P(h_1) - P(h_2)}{h_1 - h_2} \tag{5}$$

$$hf = \frac{P(f > h_2)}{P(f > 0)} \tag{6}$$

Where $P(f_0)$ represents the power at fundamental frequency, $P(h_1)$ and $P(h_2)$ are the power of first and second harmonics. The features fh1Tilt and fh2Tilt describe the two-step power dropping rate from the fundamental frequency to the second harmonic while hf specifies the cumulative power impact of those high frequency (>h_2) components.

3. Classification design

Some of the six extracted features were highly correlated, and therefore there would be redundant information if all six features were used in a classifier. Moreover too much redundancies can cause an over-fitting problem which inevitably decreases the classifier performance. Thus principle component analysis (PCA) transformed the data such that the variability of the data was accounted for by fewer (rotated) features than the original set. A Shapiro-Wilk normality tests

[15] established that a normal Gaussian model would suffice for describing the subset data distribution for each emotion and vowel type in the classifier. The average Shapiro results and overall importance of the 6 principle components is given in Table 3. The separability of five emotions can be seen in Figure 3, a scatter plot of the first two PCA components for monophthong **a**:. Different colored circles indicate different emotional entries. It can be observed that even with only two dimension the 5 emotions are well separated, there are not substantial overlaps among different clusters. In particular boredom is quite separate from anger and happiness. Similar finding were observed for the other vowels.

Table 3. *Averaged Shapiro result and overall importance of PCA components monophthongs*

	PC1	PC2	PC3	PC4	PC5	PC6
Importance	31%	25%	18%	14%	7%	5%
Shapiro w	0.47	0.43	0.19	0.22	0.52	0.45
Shapiro p	0.96	0.97	0.95	0.97	0.98	0.97

The ideal number of principle components to compactly characterize the six features is constrained by the open-test recognition performances. Multiple experiments with respect to all four sustained monophthongs demonstrated that the over-fitting problem negatively impacted on the cross-gender open-test performances when using the first five PCA components. Therefore recognition training and testing was carried out on the first four PCA components which contributed 88% of the dataset variability.

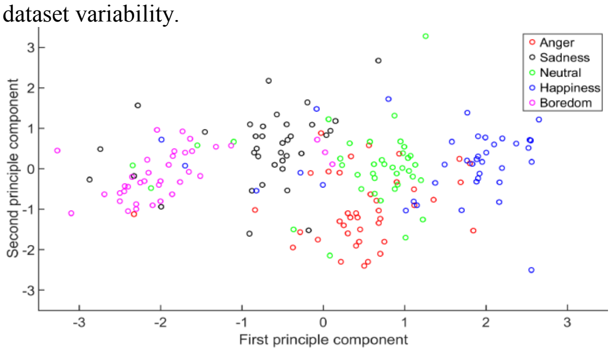


Figure 3: *PCA first two components scatter plot*

In this study, classification was conducted in three stages for different monophthongs. The first two stages were designed to classify emotions within same gender, i.e., the testing data and training data came from the same gender domain, either male or female, while the third stage was a cross-gender open-test where the testing and training data were from female and male datasets respectively (both of which had a similar sample size). A double round Robin quadratic Gaussian classification was deployed throughout.

The double round Robin classification is more powerful than those typical one-vs-all classifiers on the restricted training multi-class classification problem [16]. In this study, it turned the five-class problem into twenty binary problems, one for classifying each pair of emotional classes. For each binary problem, a base learner which was a four dimension quadratic Gaussian classifier decided to which of its two classes an input testing sample was more likely to belong. The winning emotion would be assign a “ticket”. After twenty judgments, the emotion which had accumulated the most tickets would be the final prediction of the testing entry. Possible ties would be broken by preferring the emotion with more training samples. All the testing samples would go through this procedure one by one and a confusion matrix would be formed at the end of the process.

Furthermore in order to investigate whether the low frequency speech components contributed most of emotion separation, the classification process was repeated on features calculated from low-pass filtered vowel tokens. The low-pass filter was a FIR filter with a cutoff frequency of 1000Hz.

4. Result and Discussion

Recognition confusion matrices for four monophthongs under male, female and cross-gender stages are given in Table 4. Instead of the number of samples, the percentages of correctly identified tokens, for each emotion are shown. Overall hit-rates (the sum of diagonal elements divided by the entire sum in each matrix) across all categories are given in Table 5 together with the results from the filtered data.

Results show that anger is obviously the best recognized emotion regardless of the vowel types and genders with an average hit-rate of 75% followed by boredom and neutral, although these had varying performances with respect to different categories (male **o:** the best and cross-gender **i**: the worst). In some cases, such as the cross-gender test for **i**:, sadness tokens were substantially misclassified as boredom. For the most part happiness often confused with anger, especially for the monophthong **a**: and **o**: which could be expected when looking at Figure 3. This was partly because both anger and happiness have a short intense open phase and a long relaxed closed phase. Thus it was difficult to separate those time domain features for them, however the differences from their frequency-domain could still provide meaningful classification clues. The relatively more abrupt glottal air puffs when speaking in anger raised the power of first and second harmonics in the frequency plot, resulting in less steep spectral tilt than that of happiness. This can been seen in analysis of efficacy of the time and frequency domain features in Table 6. The frequency features alone can provide much better anger and happiness recognition than the time features alone and in the case of **a**: and **o**: the recognition performances of the combination of six features even cannot compete with frequency features alone. There was also a difference between the vowels in the emotion recognition scores with rates for the back vowels **a**: and **o**: considerably better than those for the front vowel **i**: and **e**:. Male and female had almost the same average overall recognition hit-rate, 64% and 63% respectively, when the training and testing sets were of the same gender. In the cross-gender scenario (different genders for the training and testing sets) the recognition performance decreased sharply to around 50%. This is possibly due to gender differences in the glottal pulse features coupled with the insufficient cross-gender training which is bounded by the size of the applied corpus. Ideally more extensive data especially the training data should be required.

Table 5. *Overall hit-rate (%) table for raw monophthongs (bold) and filtered ones under three different stages.*

<table>
<tr><th></th><th><b>a:</b>|a:</th><th><b>e:</b>|e:</th><th><b>i:</b>|i:</th><th><b>o:</b>|o:</th><th>Average</th></tr>
<tr><td>Male</td><td><b>67</b>|66</td><td><b>61</b>|59</td><td><b>61</b>|58</td><td><b>62</b>|60</td><td><b>63</b>|61</td></tr>
<tr><td>Female</td><td><b>74</b>|67</td><td><b>58</b>|53</td><td><b>56</b>|50</td><td><b>67</b>|61</td><td><b>64</b>|58</td></tr>
<tr><td>Cross-gender</td><td><b>53</b>|44</td><td><b>52</b>|46</td><td><b>50</b>|43</td><td><b>56</b>|48</td><td><b>53</b>|45</td></tr>
<tr><td>Average</td><td><b>65</b>|59</td><td><b>57</b>|53</td><td><b>56</b>|50</td><td><b>62</b>|56</td><td></td></tr>
</table>

5. Conclusions

This work highlighted the capability of the IAIF transformed glottal pulse signals in emotion recognition using their six time and frequency related features. The whole process was free of

Table 4. *Emotion recognition hit-rate (%) confusion matrices for different monophthongs under three different stages.*

		Male					Female					Cross-gender				
		Angry	Bored	Happy	Neutral	Sad	Angry	Bored	Happy	Neutral	Sad	Angry	Bored	Happy	Neutral	Sad
a:	Angry	**90.0**	0	10.0	0	**0**	**85.0**	5.0	10.0	0	0	**75.0**	0	25.0	0	0
	Bored	0	**76.5**	5.9	0	17.6	0	**52.4**	4.8	19.0	9.5	16.7	**44.4**	11.1	22.2	5.6
	Happy	37.5	6.3	**37.5**	12.5	6.3	21.7	0	**69.6**	8.7	0	56.5	4.3	**21.7**	13.0	4.5
	Neutral	0	25.0	0	**55.0**	20.0	0	19.0	9.5	**66.7**	4.8	4.8	4.8	14.3	**61.9**	14.3
	Sad	0	16.7	0	11.1	**72.2**	0	4.5	0	9.1	**86.4**	0	0	31.8	0	**68.2**
e:	Angry	**72.2**	16.7	5.6	0	5.6	**74.0**	7.4	14.8	0	3.7	**59.3**	0	22.2	0	18.5
	Bored	0	**54.2**	0	41.7	4.2	0	**73.9**	0	21.7	4.3	0	**47.8**	13.0	17.4	21.7
	Happy	5.3	10.5	**52.6**	21.1	10.5	20.7	31.0	**41.4**	6.9	0	19.4	9.7	**38.7**	6.5	25.8
	Neutral	0	13.6	9.1	**72.7**	4.5	0	36.4	0	**63.6**	0	0	36.4	9.1	**45.5**	9.1
	Sad	4.5	13.6	4.5	27.3	**50.0**	9.5	33.3	4.3	4.8	**48.1**	4.8	14.3	0	4.8	**76.2**
i:	Angry	**70.6**	11.8	11.8	0	5.9	**90.0**	5.0	5.0	0	0	**80.0**	0	15.0	5.0	0
	Bored	0	**70.8**	0	4.2	25.0	3.8	**61.5**	3.8	7.7	23.1	11.5	**42.3**	15.4	23.1	7.7
	Happy	11.1	22.2	**38.9**	16.7	11.1	33.3	6.7	**33.3**	13.3	13.3	33.3	13.3	**46.7**	0	6.7
	Neutral	0	45.5	0	**50.0**	4.5	8.7	43.5	0	**21.7**	26.1	8.7	13.0	17.4	**60.9**	0
	Sad	0	11.8	0	11.8	**76.5**	8.7	21.7	0	0	**69.6**	4.3	43.5	21.7	8.7	**21.7**
o:	Angry	**71.4**	0	28.6	0	0	**72.7**	0	18.2	9.1	0	**81.8**	9.1	0	9.1	0
	Bored	0	**75.0**	0	0	25.0	0	**50.0**	0	37.5	12.5	0	**47.5**	12.5	27.5	12.5
	Happy	60.0	0	**30.0**	0	10.0	40.0	0	**60.0**	0	0	58.0	0	**42.0**	10.0	0
	Neutral	0	12.5	12.5	**75.0**	0	8.3	16.7	0	**58.3**	16.7	0	25.0	25.0	**50.0**	0
	Sad	0	25.0	0	12.5	**62.5**	0	0	0	14.3	**85.7**	7.1	14.3	14.3	21.4	**42.9**

manual involvement and computational hassle. Theoretically it could extend to any emotional corpus of any size with only requirements of sentence recordings and corresponding contexts. Original and low-pass filtered speech signal were both investigated. Results indicated that male and female data had decent recognition performances with average overall hit-rates of 64% and 63% respectively and the gender differences may negatively impact on the recognition performances when database were mixed with dual gender speakers such that for a more sophisticated judgement, two hierarchy structure should be built where gender identification should be on the top of emotion classification. Besides, most glottal characteristics associated with emotion discrimination was confirmed to come from the low frequency portions of the speech especially for male speakers.

In future, a more balanced and expanding corpus should be created so as to reinforce the outcome's reliability and minimize the emotion's variability within speakers and sentences. The weights of features extracted from time domain and frequency domain should be treated differently for some specific pairs of emotions like anger and happiness to increase correct recognition ratio. The capture of the nonlinear relations present in the glottal features with more complex non-linear model may also benefit to the recognition performances. Practically the fusion of linguistic and prosody characteristics will complement to glottal pulse features and as a whole push the emotion recognition to a more robust level.

Table 6. *Two-class (anger and happiness) average overall hit-rates (%) for different features and monophthongs.*

	a:	e:	i:	o:
Time features	71.1	65.0	66.4	63.5
Frequency features	77.0	75.1	69.0	71.7
All features	72.1	74.0	70.0	65.4

6. References

[1] Atal BS, "Speech analysis and synthesis by linear prediction of the speech wave", The Journal of the Acoustical Society of America. , Vol.50, p.637-655,1971

[2] Ksr Murty B Yegnanarayana, "Combining evidence from residual phase and MFCC features for speaker recognition", IEEE signal processing letters. , Vol.13(1), p.52-55,2006.

[3] E. Moore II, M. Clements, J. Peifer, and L. Weisser, "Investigating the role of glottal features in classifying clinical depression," in Proceedings of the 25th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, pp. 2849–2852, September 2003

[4] K. E. Cummings and M. A. Clements, "Analysis of the glottal excitation of emotionally styled and stressed speech," Journal of the Acoustical Society of America, vol. 98, no. 1, pp. 88–98,1995.

[5] E. Moore II, M. A. Clements, J. W. Peifer, and L. Weisser, "Critical analysis of the impact of glottal features in the classification of clinical depression in speech," IEEE Transactions on Biomedical Engineering, vol. 55, no. 1, pp. 96–107, 2008

[6] Alku, P. Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering. Speech communication, 11(2), 109-118, 1992

[7] J. Fürnkranz, "Round Robin Classification", Applied physics letters.,Vol.2(4), p.721-747, 2000.

[8] Corinna Cortes Vladimir Vapnik, "Support-Vector Networks", Machine learning. , Vol.20(3), p.273-297,1995

[9] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, and B.Weiss, "A Database of German Emotional Speech," Proc. Ninth European Conf. Speech Comm. and Technology, pp. 1517-1520, 2005.

[10] Kisler, T. and Schiel, F. and Sloetjes, H. (2012): Signal processing via web services: the use case WebMAUS, Proceedings Digital Humanities, Hamburg, Germany, Hamburg, pp. 30-34, 2012

[11] Cassidy, S. and J. Harrington, "Multi-level annotation in the Emu speech database management system", Speech Communication, 33, 61-77, 2001

[12] Bier SD, Watson CI, McCann CM. "Using the perturbation of the contact quotient of the EGG waveform to analyze age differences in adult speech", [J]. J Voice, 2014

[13] A. I. Iliev and M. S. Scordilis, "Spoken emotion recognition using glottal symmetry," EURASIP Journal on Advances in Signal Processing, Article ID 624575, pp. 1–11, 2011

[14] Proakis, J.G. and Manolakis, D.G, "Digital Signal Processing", Upper Saddle River, NJ: Prentice-Hall, pp 910–913,1996

[15] Shapiro, S.S. and Wilk, M.B, "An analysis of variance test for normality (complete samples)", Biometrika, 52, 591–611, 1965.

[16] Bo Chen Guo-Zheng Li Mingyu You, "Multi-class feature selection using Pairwise-class and All-class techniques", 2010 IEEE International Conference on Bioinformatics and Biomedicine Workshops (BIBMW), p.644-647,2010.

Can Australian English listeners learn non-native vowels via distributional learning?

Jia Hoong Ong[1,2], *Josephine Terry*[1,2], *Paola Escudero*[1,2]

[1]The MARCS Institute for Brain, Behaviour and Development, Western Sydney University
[2]Australian Research Council Centre of Excellence for the Dynamics of Language

jia.ong@westernsydney.edu.au, j.terry@westernsydney.edu.au,
paola.escudero@westernsydney.edu.au

Abstract

Inconsistent findings have been reported for distributional learning of vowels, possibly due to interference from learners' native phonological (L1) categories. Native Australian-English (AusE) listeners were exposed to unimodal and bimodal distributions of a continuum spanning Dutch /ɑ/-/a:/, which is perceived moderately well by AusE listeners. Despite sustaining learners' attention during the training phase (c.f. passive training), the distribution groups did not differ in their pre-post vowel discrimination, suggesting a lack of distributional learning. Our results imply that learners do not benefit from such rapid learning of contrasts that are perceived with high accuracy due to learners' L1 categories.

Index Terms: distributional learning, non-native speech perception, vowel perception, vowel discrimination

1. Introduction

Distributional learning refers to the process with which learners acquire knowledge by tracking frequency of occurrence of perceptual items [1]. This form of statistical learning is implicated in the acquisition of various phonological categories including consonants [2], [3], vowels [4], [5], and lexical tones [6]. For example, Hindi-learning infants will encounter speech sounds along a continuum of voice-onset time (VOT) ranging from -100ms to +20ms that can be modelled as two normal distributions with one peak at around -85ms and another at around +13ms [7]. Conversely, English-learning infants will encounter speech sounds along the same continuum that can be modelled as a normal distribution with a 0ms peak. This difference in the distributional peaks along that range of VOT—in particular, a bimodal or two-peak distribution for Hindi learners and a unimodal or single-peak distribution for the English learners—may explain the formation of two voicing categories by Hindi listeners (prevoiced and voiceless) and the formation of only one voicing category by English listeners (voiceless).

Distributional learning has been demonstrated in the laboratory by systematically manipulating the frequency of training tokens from a continuum spanning a contrast (e.g., /d/-/t/) presented to learners. Learners who were exposed to relatively more prototypical tokens of the contrast (i.e., those who were exposed to a bimodal distribution) were more likely to show greater discrimination of that contrast after training whereas those who were exposed to relatively more ambiguous tokens of that contrast (i.e., those exposed to a unimodal distribution) were less likely to show any improvement in discrimination after training [1], [3], [4], [8]. Research suggests that infant- and adult-learners are able to learn phonological categories distributionally and so distributional learning may play a crucial role in first-language (L1) and second-language (L2) acquisition [9]–[11]—though some argue that distributional learning may not be as effective for adults than for infants [12].

While distributional learning appears to be used by all language learners, its limits are still unclear. Limits may be related to individual and population differences or the learning environment in acquiring phonological categories via distributional learning. For instance, distributional learning appears to be influenced by whether an attentive or passive learning environment was used; the former appears to be more effective for distributional learning [6]. Concerning individual differences, distributional learning may not be equally effective for all learners as there may be individual differences in the perception of target speech sounds. [13]–[15]. This paper examines a possible factor that contributes to such perceptual differences; one that may arise from listeners' native language. It is widely accepted in non-native speech perception research that a particular non-native contrast may be perceived accurately or poorly by non-native listeners depending on their native phonological categories [10], [11], [16]–[18].

In this paper, we investigate whether native listeners of a particular language (Australian English, AusE) may be disadvantaged in learning a L2 contrast (Dutch /ɑ/ and /a:/) via distributional learning. Previous studies on distributional learning of Dutch /ɑ/-/a:/ have been successful when conducted with native Spanish listeners (i.e., improvement in discriminating that contrast by those in the bimodal condition but not those in the control condition in which participants listened to music; [4]) but not with native AusE listeners (i.e., both bimodal and unimodal conditions improve after training; [19]). These diverging results may be due to listeners' initial perception of the Dutch contrast. That is, Spanish listeners perceive the Dutch /ɑ/-/a:/ contrast as a single category, while AusE listeners perceive it as belonging predominantly to two categories [20]. This difference in categorisation across population leads to different degrees of discrimination accuracy: Spanish listeners will have poor discrimination performance while AusE listeners will have moderate to good discrimination performance (better than native Spanish listeners but not native-like as would be the case with native Dutch listeners). Indeed, a direct comparison of initial performance of this Dutch vowel contrast by native Spanish listeners (data from [8], [21]) and native AusE listeners (data from [19]) reveals that native AusE listeners have significantly higher accuracy than native Spanish listeners ($t(220)= 7.336, p< .001$).

We propose that distributional learning may be constrained by learners' initial perception of the target contrast: the mechanism may most benefit those that are poor perceivers of a contrast (i.e., those that perceive the contrast to be of a single category) but it may not be effective for listeners who are already relatively good perceivers of that contrast (i.e., those that perceive the contrast to belong to two separate categories generally). However, in order to conclude such, it is necessary to examine whether the lack of distributional learning of Dutch /ɑ/-/a:/ by AusE listeners is due to the learning environment. Thus, in this study, we presented the same number of training tokens as that presented to native Spanish listeners and to native AusE listeners in previous distributional learning studies [4], [19]. However, unlike those studies, we included an auditory vigilance task during the training phase to sustain learners' attention to training tokens, which has been shown to be more effective in eliciting distributional learning [6].

If distributional learning is constrained by how learners initially categorise the target contrast, then, we predict no distributional learning by native AusE listeners (i.e., both bimodal and unimodal distribution conditions improve in their discrimination of Dutch /ɑ/ and /a:/ after training). If this hypothesis is supported *despite* sustaining learners' attention to training tokens, then this can be taken as strong evidence of the role of initial perception of target contrast in distributional learning, particularly when the results of this experiment are considered with previous distributional learning studies with native Spanish listeners.

2. Method

2.1. Participants

Participants were 51 (28 females, 23 males) native AusE-listening undergraduates (Age range= 17-59, *M*= 22.45, *SD*= 6.44) from Western Sydney University. Forty-two participants spoke one or more languages other than English although none had prior exposure to Dutch. Participants were randomly allocated to one of two distribution conditions: Unimodal (*n*= 25) and Bimodal (*n*= 26). The average age of participants and the ratio of monolinguals to bilinguals were equivalent across distribution conditions (Age, *t*(49)= 0.617, *p*= .540; Monolinguals, $\chi^2(1)$= 0.187, *p*= .666). All provided their informed consent prior to participating and they were given course credit in return for participating.

2.2. Stimuli

2.2.1. Test stimuli

Test stimuli for this experiment consisted of naturalistic and synthetic tokens of Dutch vowels /ɑ/ and /a:/. The naturalistic tokens, which were selected from the corpus of Adank et al [22], were naturally produced Dutch vowels /ɑ/ and /a:/, taken from a /s-V-s/ syllable embedded in a carrier sentence. Sentences were produced by 10 male and 10 female speakers of Standard Northern Dutch, resulting in 20 tokens of /ɑ/ and 20 tokens of /a:/. The synthetic tokens were produced in Praat [23] and constituted good examples of /ɑ/ and /a:/ [21], [24]. They were 140ms in duration and the fundamental frequency (F0) fell between 150 and 100 Hz, representing a male voice.

2.2.2. Training stimuli

Training stimuli consisted of eight vowel training tokens that fall along the /ɑ/ to /a:/ continuum, produced in the same manner as the synthetic test stimuli. The training tokens differed from one another only in terms of F1, F2, and F3 spectral values such that from one end of the continuum to the other, the vowel morphs from an /ɑ/ (Token 1) to an /a:/ (Token 8).

2.3. Equipment

The experiment was presented using MATLAB 2012b [25] on an Acer TravelMate P653 laptop that used Windows 7 as the operating system. The auditory stimuli were presented at a comfortable volume using a pair of Sennheiser HD650 headphones connected to an Edirol USB Audio Capture UA-25EX audio interface.

2.4. Procedure

There were three phases to the experiment: pre-test discrimination, training, and post-test discrimination.

2.4.1. Pre- and Post-test discrimination

During pre-test and post-test, participants were required to complete a two-alternate forced choice discrimination task (XAB), in which participants determined whether the first sound (X) was similar to the second (A) or third sound (B). In each trial, the ISI was 1.2s. There were 84 trials in each test phase. The A and B tokens were always the synthetic test vowels in all trials, with the position of the synthetic /ɑ/ to /a:/ equally distributed across all trials as the second or third sound. The X tokens, on the other hand, were sampled from the naturally-produced test stimuli in 80 trials and in the remaining four trials, the X tokens were the synthetically-produced test-stimuli (these four trials constituted the 'catch trials'). The /ɑ/ to /a:/ vowels were X tokens an equal number of times in each test phase. Participants could only proceed to the next trial once a response had been collected via a mouse click of a button on the screen. Breaks were provided throughout each test phase.

2.4.2. Training

During training, participants were exposed to the training tokens. Participants in the two distribution conditions heard the same total number of training tokens (i.e., 128 tokens) but the conditions differed in the distribution of the training tokens. Specifically, participants in the Unimodal condition heard Tokens 4 and 5 (i.e., ambiguous /ɑ/ to /a:/ vowels) the most frequently whereas those in the Bimodal condition heard Tokens 2 and 7 (i.e., prototypical /ɑ/ to /a:/ vowels) the most frequently (see Table 1).

Table 1. *Frequency distribution of training tokens in Unimodal and Bimodal conditions*

Training Token	1	2	3	4	5	6	7	8
Unimodal	8	8	16	32	32	16	8	8
Bimodal	8	32	16	8	8	16	32	8

Following the training protocol of a previous distributional learning study [6] and as an extension of a previous distributional learning study involving the same vowel contrast [19], participants were required to completed an auditory vigilance task during training, which ensures attentive listening. Participants were instructed during the training phase that they were to listen to the sounds attentively and that some of the sounds would be beeps (sine wave tones). When they heard a beep, they were required to circle the sound number that corresponded to the beep on a paper response sheet. A total of 16 beeps were interspersed randomly within the 128 training tokens. The entire training phase took approximately 3mins to complete.

3. Results

We first determined whether the two distribution conditions differed in their Pre-test scores. The Pre-test scores of the two distribution conditions were not significantly different, either when the Pre-test scores were considered for each vowel (/ɑ/, *t*(49)= 0.270, *p*= .788; /a:/, *t*(49)= 0.552, *p*= .584) or as a whole (*t*(49)= .496, *p*= .622), suggesting that any difference at Post-test between the two distribution conditions may be attributed to the training itself.

A Mixed ANOVA with a between-subject factor Distribution Condition (Unimodal vs. Bimodal) and within-subject factors Session (Pre-test vs. Post-test) and Vowel (/ɑ/ vs. /a:/) revealed a main effect of Session (*F*(1, 49)= 38.069, *p*< .001, η_p^2= 0.437) and a significant Session by Vowel interaction (*F*(1, 49)= 11.486, *p*= .001, η_p^2= 0.190). Simple main-effect analysis revealed that, from Pre- to Post-test, participants showed greater improvement in discriminating /a:/ (Pre-test: *M*= 0.661, *SE*= .023; Post-test: *M*= 0.764, *SE*= .023) than /ɑ/ (Pre-test: *M*= 0.693, *SE*= .018; Post-test: *M*= 0.732, *SE*= .020). Importantly, the ANOVA revealed no significant Session by Distribution Condition interaction, which suggests that the two distribution conditions did not differ in their discrimination accuracy from Pre- to Post-test (see Figure 1).

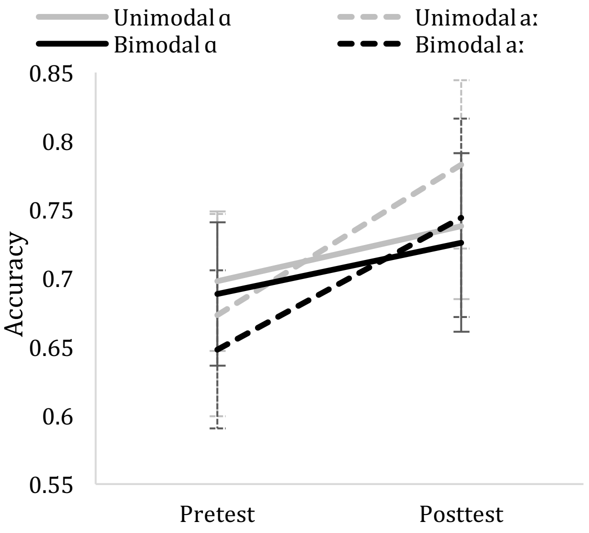


Figure 1: *Accuracy scores from Pre- to Post-test for Unimodal (gray) and Bimodal (black) distribution conditions by vowel (*ɑ *in solid and* a: *in dashed line)*

4. Discussion

Results of this study indicate that despite sustaining learners' attention to the training tokens, native AusE learners do *not* show distributional learning of Dutch /ɑ/-/a:/ vowels; both learners in the bimodal and unimodal distribution conditions improved in their discrimination of /ɑ/ and /a:/ after training. The improvement seen by both distribution conditions is likely due to the fact that participants performed the same test twice (practice effects). Unexpectedly, across both distribution conditions, learners showed greater pre-post improvement in discriminating /a:/ than /ɑ/. This larger improvement for a more peripheral vowel (Dutch /a:/) is reminiscent of the natural referent vowel framework [26], in which it is argued that listeners have an underlying perceptual bias for vowels in the peripheral of F1/F2 vowel space. Though it is unclear how distributional learning may contribute to such asymmetry, our finding on asymmetric improvement suggests a change in perception of each Dutch vowel by AusE listeners. Further work is needed to investigate this finding.

With respect to the lack of distributional learning seen among AusE listeners in this study, the results complement a similar previous study that used a passive training paradigm [19], suggesting that the learning environment (i.e., attentive or passive) does not impact their learning of Dutch /ɑ/-/a:/. It may be the case that on top of an attentive listening paradigm, the stimuli need to be exaggerated or enhanced for such learning to be successful [8]. Enhanced stimuli are argued to draw learners' attention to the acoustic differences of the target contrast [12] and so a larger gain in distributional learning may be observed. However, it may also be the case that since AusE listeners are already good perceivers of that contrast, exaggerating the stimuli will not yield any advantage. Work is currently underway in our laboratory to investigate whether enhanced bimodal distributions will lead to distributional learning in this population.

Native Spanish listeners, on the other hand, have been successful in demonstrating distributional learning of Dutch /ɑ/-/a:/ even when trained passively [4], [8]. Thus, we propose that this population difference in distributional learning of Dutch /ɑ/-/a:/ may be due to learners' L1 influence. Specifically, it appears that distributional learning is affected by how a L2 contrast is categorised in terms of learners' L1 phonological categories. When a L2 contrast is perceived as a single category—which leads to poor discrimination performance, as is the case with native Spanish listeners on the Dutch /ɑ/-/a:/ vowels—distributional learning appears to be effective. However, when the L2 contrast is already perceived as two different categories (although not to the extent of native listeners)—as is the case with native AusE listeners on the Dutch /ɑ/-/a:/ vowels—then learners do not appear to benefit from distributional learning.

Two caveats are worth mentioning. As demonstrated in the present study and in a previous study [19], native AusE listeners do not show distributional learning of Dutch /ɑ/-/a:/ vowels when the standard amount of training tokens as used in most distributional learning studies (i.e., 128 training tokens) is presented. It may be the case that listeners with relatively good perception of the target contrast may need longer training or more training tokens in order to learn the target contrast distributionally. Future studies can test this prediction by replicating the present experiment with double the amount of training [27]. Secondly, it could also be the case that our behavioural measure was not sensitive enough to register any learning by native AusE listeners after such rapid training. To investigate this possibility, future research using electroencephalography (EEG) could shed light on whether such rapid learning can occur preattentively using EEG measures such as mismatch negativity (MMN).

To further support our proposal, future studies could compare distributional learning of the Dutch /ɪ/-/i/ contrast by native AusE and native Spanish listeners. The categorisation pattern of the Dutch /ɪ/-/i/ contrast by native AusE and Spanish listeners is the reverse of the Dutch /ɑ/-/a:/: the Dutch /ɪ/-/i/ is perceived as a single category by AusE listeners but essentially as two categories by Spanish listeners [20]. Consequently, AusE listeners are reported to be poor perceivers of the Dutch /ɪ/-/i/ contrast whereas Spanish listeners are relatively good perceivers of the same Dutch contrast although not to the extent of native Dutch listeners [20]. Thus, if our proposal is true, that is, if distributional learning is influenced by how accurately learners' perceive the L2 contrast prior to distributional training, then we may expect AusE listeners, but not Spanish listeners, to show distributional learning of Dutch /ɪ/-/i/ contrast. Specifically, AusE listeners in the bimodal condition, but not those in the unimodal condition, should improve in their discrimination of Dutch /ɪ/-/i/ after training, whereas Spanish listeners should improve regardless of distribution condition due to practice effects from performing the test twice.

5. Conclusion

In conclusion, this study demonstrated that the lack of distributional learning of Dutch /ɑ/-/a:/ among AusE listeners is not due to the learning environment (i.e., attentive listening). When this study is considered with previous studies on distributional learning of Dutch /ɑ/-/a:/ by native AusE and native Spanish listeners, we propose that how L2 contrasts are categorised in learners' L1 can predict whether learners show distributional learning. Specifically, if learners perceive the L2 contrast as a single category (and therefore are poor perceivers of that contrast in the case of Spanish listeners), distributional learning appears to be effective. On the other hand, if learners perceive the L2 contrast as two separate categories (and thus are good perceivers of that contrast), they may not benefit from distributional learning—at least not with natural bimodal distributions with 128 training tokens. Further work is needed to examine this proposal more closely in order to broaden our understanding of the limits of distributional learning in acquiring L2 speech sounds.

6. Acknowledgements

We would like to thank Christina Quattropani for help with data collection. This research was supported by the Australian Research Council Centre of Excellence for the Dynamics of Language research grant (project number CE140100041) awarded to PE.

7. References

[1] J. Maye, J. F. Werker, and L. Gerken, "Infant sensitivity to distributional information can affect phonetic discrimination," *Cognition*, vol. 82, no. 3, pp. B101–B111, 2002.

[2] J. Maye, D. J. Weiss, and R. N. Aslin, "Statistical phonetic learning in infants: Facilitation and feature generalization," *Dev. Sci.*, vol. 11, no. 1, pp. 122–134, 2008.

[3] J. Maye and L. Gerken, "Learning phonemes without minimal pairs," in *Proceedings of the 24th Annual Boston University Conference on Language Development*, 2000, vol. 2, pp. 522–533.

[4] P. Escudero and D. Williams, "Distributional learning has immediate and long-lasting effects," *Cognition*, vol. 133, no. 2, pp. 408–13, Nov. 2014.

[5] K. Wanrooij, P. Boersma, and T. L. van Zuijen, "Fast phonetic learning occurs already in 2-to-3-month old infants: An ERP study," *Front. Psychol.*, vol. 5, no. FEB, 2014.

[6] J. H. Ong, D. Burnham, and P. Escudero, "Distributional learning of lexical tones: A comparison of attended vs. unattended listening," *PLoS One*, vol. 10, no. 7, p. e0133446, 2015.

[7] L. Lisker and A. S. Abramson, "A cross-language study of voicing in initial stops: Acoustical measurements," *Word*, vol. 20, no. 3, pp. 384–422, 1964.

[8] P. Escudero, T. Benders, and K. Wanrooij, "Enhanced bimodal distributions facilitate the learning of second language vowels.," *J. Acoust. Soc. Am.*, vol. 130, no. 4, pp. EL206–12, Oct. 2011.

[9] J. F. Werker, H. H. Yeung, and K. A. Yoshida, "How do infants become experts at native-speech perception?," *Curr. Dir. Psychol. Sci.*, vol. 21, no. 4, pp. 221–226, Jul. 2012.

[10] P. Escudero, "Linguistic perception and second language acquisition: Explaining the attainment of optimal phonological categorization," Utrecht University, 2005.

[11] J.-W. van Leussen and P. Escudero, "Learning to perceive and recognize a second language: The L2LP model revised," *Front. Psychol.*, vol. 6, pp. 1–12, 2015.

[12] K. Wanrooij, P. Boersma, and T. L. van Zuijen, "Distributional vowel training is less effective for adults than for infants. A study using the mismatch response," *PLoS One*, vol. 9, no. 10, 2014.

[13] R. Frost, B. C. Armstrong, N. Siegelman, and M. H. Christiansen, "Domain generality versus modality specificity: The paradox of statistical learning," *Trends Cogn. Sci.*, vol. 19, no. 3, pp. 117–125, 2015.

[14] A. Cristià, G. L. McGuire, A. Seidl, and A. L. Francis, "Effects of the distribution of acoustic cues on infants' perception of sibilants," *J. Phon.*, vol. 39, no. 3, pp. 388–402, 2011.

[15] K. Wanrooij, P. Escudero, and M. E. J. Raijmakers, "What do listeners learn from exposure to a vowel distribution? An analysis of listening strategies in distributional learning," *J. Phon.*, vol. 41, no. 5, pp. 307–319, Sep. 2013.

[16] C. T. Best and M. D. Tyler, "Nonnative and second-language speech perception," in *Language experience in second language speech learning: In honor of James Emil Flege*, M. J. Munro and O.-S. Bohn, Eds. Amsterdam: John Benjamins, 2007, pp. 13–34.

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

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

[19] J. Terry, J. H. Ong, and P. Escudero, "Passive distributional learning of non-native vowel contrasts does not work for all listeners," in *Proceedings of the 18th International Congress of Phonetic Sciences (ICPhS)*, 2015.

[20] S. Alispahic, P. Escudero, and K. E. Mulak, "More vowels are not always better: Australian English and Peruvian Spanish learners ' comparable perception of Dutch vowels," in *Proceedings of the 39th annual Boston University Conference on Language Development*, 2015, pp. 40–51.

[21] P. Escudero and K. Wanrooij, "The effect of L1 orthography on non-native vowel perception," *Lang. Speech*, vol. 53, no. 3, pp. 343–365, 2010.

[22] P. Adank, R. van Hout, and R. Smits, "An acoustic description of the vowels of Northern and Southern Standard Dutch," *J. Acoust. Soc. Am.*, vol. 116, no. 3, pp. 1729–1738, 2004.

[23] P. Boersma and D. Weenink, "Praat: Doing phonetics by computer." 2013.

[24] L. C. W. Pols, H. R. C. Tromp, and R. Plomp, "Frequency analysis of Dutch vowels from 50 male speakers," *J. Acoust. Soc. Am.*, vol. 53, pp. 1093–1101, 1973.

[25] "MATLAB Release 2012b." Natick, Massachusetts, United States, 2012.

[26] L. Polka and O. S. Bohn, "Natural Referent Vowel (NRV) framework: An emerging view of early phonetic development," *J. Phon.*, vol. 39, no. 4, pp. 467–478, 2011.

[27] K. A. Yoshida, F. Pons, J. Maye, and J. F. Werker, "Distributional phonetic learning at 10 months of age," *Infancy*, vol. 15, no. 4, pp. 420–433, 2010.

The relationship between Australian English speakers' non-native perception and production of Brazilian Portuguese vowels

Jaydene Elvin[1,2], Paola Escudero[1,2], Daniel Williams[1,2,4] and Catherine T. Best[1,2,3]

[1]The MARCS Institute , Western Sydney University, Australia
[2]ARC Centre of Excellence for the Dynamics of Language, Australia
[3]School of Humanities and Communication Arts, Western Sydney University, Australia
[4]Linguistics Department, Area of Excellence – Cognitive Sciences, University of Potsdam, Germany

{j.elvin, paola.escudero, c.best}@westernsydney.edu.au, daniel.williams@uni-potsdam.de

Abstract

This study investigates the relationship between non-native perception and production of Brazilian Portuguese (BP) vowels by six Australian English (AusE) monolinguals. Participants' non-native categorization and discrimination patterns were used to predict performance in non-native production. We further investigated the acoustic similarity between participants' non-native and native vowel productions. The findings indicate a perception-production link. In particular, non-native vowel production was acoustically similar to native vowel production. We also found that perceptually difficult non-native vowel contrasts were predominantly produced as one single non-native category.

Index Terms: non-native perception, non-native production, interrelation

1. Introduction

One of the main goals of a second language (L2) learner is to be able to speak the L2 in a native-like manner. Unfortunately, for many, this goal is not achieved [1]. This is because factors such as the age of L2 acquisition, length of residence in an L2 environment, language experience and L1 vs. L2 use [2] all contribute to the degree of foreign-accented speech. Common to all of the factors is the influence of an individual's native language (L1) on their L2 pronunciation.

Problems in L2 pronunciation are thought to be caused by their L2 perceptual difficulties. Models of L2 speech perception such as the Speech Learning Model [SLM; 3], the Perceptual Assimilation Model [PAM; 4] and the Second Language Linguistic Perception model [L2LP; 5,6] all claim that L2 speech sounds are filtered by listeners' L1. For example, two non-native sounds that are perceived as one single category are known as single-category assimilation in PAM and as the NEW scenario in L2LP. A scenario of this type has been shown to result in perceptual discrimination difficulties, whereas no discrimination difficulties are found for scenarios where two non-native sounds are mapped on to two separate native categories (two-category assimilation in PAM and the SIMILAR scenario in L2LP). When two sounds in a non-native contrast are mapped on to two or more native categories (uncategorised assimilation in PAM and the SUBSET scenario in L2LP) discrimination difficulty may occur if these non-native vowels are mapped to the same overlapping native categories.

While the above scenarios predict difficulty in non-native or L2 perception, the L2LP model posits that these learning scenarios will also influence an individual's production of non-native or L2 vowels. Although the L2LP model is not the only model to suggest this (e.g. SLM investigates the limitations of a learner's ability to perceive and produce native-like sounds due to experience and age-related limitations), it is the only model to account for a relationship between perception, spoken word recognition and production at all stages of development. According to the L2LP theoretical framework, at the onset of learning, L2 production will closely match the acoustic properties of sounds produced in the speaker's L1. L2 pronunciation develops as learners adjust their perceptual mappings to match those of the L2 with the help of their lexical (word) representations.

Studies investigating the perception-production relationship have produced mixed results. A number of studies [1, 6] have indeed identified a link between perception and production and suggest that native language perceptual patterns do seem to influence L2 production patterns. However, there is a long-standing debate regarding the perception-production link which is unresolved due to mixed empirical results and the problematic nature of testing this link [7]. In particular, there are a number of methodological reasons that make the investigation of the interrelation difficult. For example, there are different task demands with different techniques of analysis. Additionally, previous interrelation studies have tended to focus on groups rather than individuals [6]. In order to appropriately investigate the relationship between perception and production, [7] suggest that data should ideally be collected from the same participants performing both tasks.

The aim of the present study is to investigate the relationship between the perception and production of non-native Brazilian Portuguese (BP) vowels by Australian English (AusE) monolinguals. In particular, we will investigate whether our participants' non-native categorization patterns and discrimination difficulties influence the way in which these same vowels are produced in the non-native production task. We control for the methodological issues suggested by [7] by using the same participants and the same stimuli across all tasks, as well as collecting native vowel production data to compare with their non-native vowel productions.

The L2LP theoretical framework is the most applicable to the present study as it specifically accounts for learners, such as our AusE participants who have no prior experience with the target language. If the findings are consistent with the L2LP theoretical framework, we would expect that our participants should produce non-native BP vowels similarly to the productions of their own closest L1 vowel categories and that participants will have difficulties producing vowels that are similar to two separate BP categories for those contrasts that are difficult to discriminate.

2. Method

2.1. Participants

This paper reports a subset of six AusE monolinguals (3 male) from [8] which is a study that investigated the non-native perception of BP vowels. They were all Australian English speakers born and raised in Western Sydney and aged between 18 and 30. The AusE participants reported little to no knowledge of any foreign language and provided informed consent.

2.2. Stimuli and procedure

We first recorded participants' productions of the 13 AusE vowels, namely, /iː, ɪ, ɪə, e, eː, ɜː, ɐ, ɐː, æ, oː, ɔ, ʊ and ʉː/ produced in the fVf context. The auditory stimuli were the same for the non-native categorization, discrimination and repetition tasks and consisted of naturally produced BP pseudo-words in the fVfe context. Target words were produced by five male and five female speakers from São Paulo, selected from the [9] corpus. The vowel in the first syllable was always stressed and corresponded to one of the seven Portuguese target vowels /i, e, ɛ, a, o, ɔ, u/. The perception and production tasks were counterbalanced. In perception, participants first completed a 2 alternate forced choice task in an XAB format, followed by a non-native categorization task. A non-native repetition task was used to elicit production data as it was the most appropriate task for monolingual participants with no experience with Brazilian Portuguese.

2.3. Data analysis for native and non-native vowel production

WebMaus [10], an online tool for automatically segmenting and labelling speech sounds was used to segment both the native and non-native vowel productions. The automatically generated start and end boundaries were checked and manually adjusted to ensure accuracy. Vowel duration was measured as the time (ms) between these start and end boundaries. Formant measurements for each vowel token were extracted at three time points (25%, 50%, 75%) following the optimal ceiling method reported in [9]. In the optimal ceiling method, for every vowel, per speaker, the "optimal ceiling" is chosen as the one that yields the least amount of variation for both the first and second formant values within the set number of annotated tokens for the vowel. Formant ceilings ranged between 4500 and 6500 Hz for females and 4000 and 6000 for males.

3. Results

3.1. Non-native perception

3.1.1. Non-native categorization

Table 1 shows the results from the non-native categorization task. Both BP /i/ and /e/ were perceived as AusE /iː/ and /ɪ/, with BP /e/ also being categorized as AusE /ɪə/ and /eː/. BP /ɛ/ was mostly classified as AusE /eː/, with some tokens also being categorized as AusE /æ/. Categorization for BP /a/ was split between AusE /æ/ as well as /ɐː/. Most tokens of BP /o/ were perceived as AusE /oː/, with a smaller amount also being classified as AusE /ʊ/. Likewise, BP /ɔ/ was generally perceived as AusE /oː/ and finally, BP /u/ was predominately categorized as AusE /ʊ/. In regards to predicting difficulty in discrimination and non-native production based on these categorization patterns, we would expect that BP /a-ɔ/ and /a-ɛ/ should be easier to discriminate and produce than the remaining contrasts as the target vowels in these contrasts were mapped to different native categories resulting in little to no perceptual overlap. The non-native categorization patterns further suggest that participants may have difficulties discriminating and producing BP /i/-/e/ and /o/-/u/ as both vowels are mapped to the same native categories causing a large amount of perceptual overlap. Additionally, the vowels in the BP contrasts, /e/- /ɛ/ and /o/- /ɔ/ are mapped to some of the same native categories and this may make these contrasts difficult to discriminate and produce.

Table 1: *Classification percentages for the six AusE listeners. The native vowel category with the highest classification percentage appears in bold and those percentages below chance (i.e. 8%) appear in grey.*

BP	AuE vowels												
	iː	ɪ	ɪə	e	eː	ɜː	æ	ɐː	ɐ	ɔ	oː	ʊ	ʉː
i	**48**	38	7	5				2					
e	**27**	13	22	5	**27**		3	3					
ɛ	7	2	5	3	**53**	2	20	8					
a			2		3	3	**43**	**43**			5		
o						7			8	7	**62**	15	2
ɔ					3	10		2		3	**75**	2	5
u			2			8			15	2	7	**45**	22

3.1.2. Non-native discrimination

Figure 1 shows the average accuracy scores for each of the six BP contrasts.

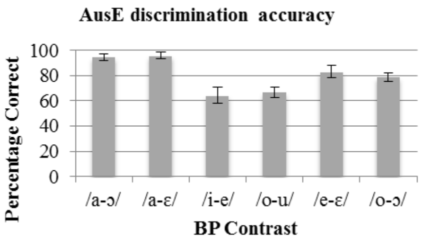


Figure 1: *Discrimination accuracy across the six BP contrasts*

The discrimination results indicate that AusE listeners had higher performance on BP /a-ɔ/ and /a-ɛ/, which is unsurprising given that there was little to no perceptual overlap for these contrasts. We would therefore predict that participants should be able to produce two distinct vowels that are similar to two different BP vowels. Furthermore, in line with the non-native categorization predictions, participants did indeed have overall lower discrimination accuracy scores for BP /i/-/e/ and /o/-/u/ and we would therefore predict that participants may produce both vowels in each contrast as the same non-native vowel category. Additionally, BP /e/-/ɛ/ and /o/-/ɔ/ were not the easiest to discriminate, however they were also not the hardest to discriminate. It is possible that participants may still have difficulty producing these vowels given perceptual overlap identified in non-native categorization, but perhaps they may not be as unstable as those contrasts with the lowest discrimination accuracy.

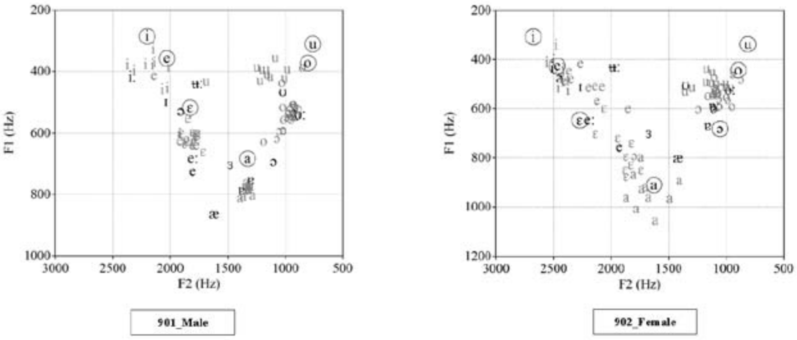


Figure 2: *A sample of one male and one female participants' F1 and F2 values for all tokens of their non-native productions of the BP vowels (grey) are displayed against the mean values for their native AusE vowel productions (black), and for the target BP vowels (black, with circles).*

3.2. Non-native production

3.2.1. Acoustic similarity between non-native vowel productions and native vowel production

Figure 2 shows an example of the F1 and F2 values for one male and one female participant's non-native production of the 7 BP vowels, together with the averaged F1 and F2 values of their own native vowel productions and the target BP vowels. We show these example vowel plots to demonstrate how non-native production of the BP vowels varies across participants and that they have not yet formed stable vowel categories for each vowel. In fact, it appears that at the initial stage of learning, non-native vowel productions are indeed acoustically more similar to their own L1 vowel categories.

Table 2: *Predicted group membership for participants' non-native vowel productions tested on AusE vowel productions. The native vowel category with the highest predicted probability appears in bold and those probabilities below chance (i.e. 8%) appear in grey.*

BP	**AusE vowels**												
	iː	**ɪ**	**ɪə**	**e**	**eː**	**ɜː**	**æ**	**ɐː**	**ɐ**	**ɔ**	**oː**	**ʊ**	**ʉː**
i	**47**	26	10	1							2		14
e	**24**	18	11	19	6	2							20
ɛ	8	8	7	**24**	23	12	6	1					15
a				2		4	**40**	17	27	8	1	2	
o						2		1	4	23	29	**40**	
ɔ						6	2	4	8	**29**	**29**	23	
u				1		1			1	11	10	**74**	2

Table 2 shows the results of a cross-language discriminant analysis conducted to determine acoustic similarity between participants' non-native productions of BP vowels and their own native vowels. We trained the model on the participants' own native vowel productions and tested it against their non-native productions of BP using F1, F2, F3 and duration as input parameters. Instead of reporting the percentage of times a BP vowel was categorized as a native, we report the probabilities of group membership averaged across the BP vowel tokens. The benefit of reporting probabilities in the present study is that it takes into account that some BP tokens may be acoustically close to more than one vowel, which can be masked by categorization percentages. The results indicate that participants' productions of BP vowels at the initial state are indeed acoustically similar to their own native vowel categories. In particular, BP /i/ and /e/ are acoustically similar to AusE /iː/, /ɪ/, /ɪə/ and /ʉː/. Participants produced BP /ɛ/ as acoustically similar to AusE /e/, /eː/ and /ʉː/. The non-native productions of BP /a/ are acoustically similar to /æ/, /ɐː/ and /ɐ/. We also found that participants produced BP /ɔ/ with equal chance that the vowel would be classified as AusE /ɔ/ and /oː/ with 23% chance the tokens would be classified as AusE /ʊ/. Furthermore, there was a 40% chance that the non-native productions of BP /o/ would be classified as AusE /ʊ/ (with 29% chance for AusE /oː/ and 23% chance for AusE/ɔ/). Additionally, there was a 75% chance that the productions for BP /u/ would be classified as AusE /ʊ/ with an 11% chance for produced similarly to AusE /ɔ/ and 10% for AusE /oː/.

3.2.2. Acoustic similarity between non-native BP vowel production and target BP stimuli

Table 3: *The percentage of non-native vowel tokens classified as the intended BP vowel category. The highest classification percentage appears in bold.*

Non-native	**BP vowel category**						
↓	**a**	**e**	**ɛ**	**i**	**o**	**ɔ**	**u**
a	**80**					20	
e		**49**	46	5			
ɛ	11	14	**70**	2		4	
i		**63**	3	32		2	
o			2		**66**	28	5
ɔ	5		3		32	**60**	
u			8		**67**	2	23

To determine acoustic similarity between the target BP tokens and non-native productions we ran an additional cross-language discriminant analysis with the model trained on target BP vowel tokens and tested on our participants' non-native vowel productions, using the same input parameters from the previous discriminant analysis. The model yielded 85.7% correct classification for the trained BP vowels and 54.1% correct classification for the non-native vowel productions. This suggests that only half of the non-native BP vowel productions were produced close to the target BP vowels. Table 3 shows the percentage of times that each non-native vowel production was correctly classified as the target BP token.

As predicted it seems that participants were able to produce vowels that were acoustically similar to two different vowels for the BP contrasts /a/-/ɛ/ and /a/-/ɔ/ which were perceptually easy to discriminate. There was very little acoustic overlap between the non-native productions of BP /a/ and /ɛ/, with BP /a/ being correctly classified for 80% of the non-native vowel productions and 70% correct classification for BP /ɛ/. Likewise, there is no overlap between BP /a/ and /c/ with 60% correct classification for BP /ɔ/.

We further find that the non-native productions of the BP contrasts with lower discrimination accuracy scores, namely BP /i/-/e/ and /o/-/u/ were indeed less stable and more varied as participants seem to be unsuccessful in producing non-native vowels that were acoustically similar to two separate BP vowel categories. For example, in the BP /i/-/e/ contrast only 32% of the BP /i/ productions were correctly classified as the majority of tokens were classified as BP/e/. Likewise, only 23% of the productions of the BP /u/ vowel were correctly classified, with 67% of the non-native productions of BP /u/ classified as BP /o/. The AusE participants' productions of the vowels in the BP /e/-/ɛ/ and BP /o/-/ɔ/ contrasts was also less stable and more varied. While 49% of the non-native vowel productions of BP /e/ were correctly classified, 46% were incorrectly classified as BP /ɛ/. For BP /o/- /ɔ/, 66% of the non-native vowel productions of BP /o/ were correctly classified, with 28% incorrectly classified as /ɔ/. Furthermore, 60% of the non-native productions of BP /ɔ/ were correctly classified, with 32% being incorrectly classified as BP /o/.

4. Discussion

The aim of the present study was to investigate the relationship between the perception and production of non-native Brazilian Portuguese (BP) vowels by Australian English (AusE) monolinguals. The L2LP theoretical framework posits that learners will initially perceive and produce vowels of the target language in the same manner in which they perceive and produce vowels in their own native language. Our findings support this claim as we found that our participants' perception of BP was influenced by their native language and produced BP vowels similar to their own native categories.

Our findings also support the L2LP claim that perception is linked to and perhaps precedes production. In particular, we found that the non-native contrasts which are easy to perceive, the two vowels in that contrast were produced as similar to two separate BP categories. We also found that non-native vowel production was less stable and more varied for those vowel contrasts which had lower overall discrimination accuracy. This finding suggests that not only do learners struggle to perceive a difference between the two vowels in the BP contrast, but they also find it difficult to produce these vowels as separate categories. Given the particularly low accuracy scores in the discriminant analysis for BP /i/ and /u/ we would expect that speakers will likely have a heavy accent when producing these vowels. We further found that non-native vowels produced in the BP /e/-/ɛ/ and /o/-/ɔ/ contrasts were less stable and more varied, even though these contrasts were not the most difficult to discriminate. It is highly likely that is because their non-native categorization patterns show a considerable amount of acoustic overlap.

While our findings are in-line with the L2LP model, our findings are also consistent with predictions found in the PAM framework as listeners' non-native categorization patterns predicted difficulty in non-native discrimination. Although PAM does not explicitly account for non-native production, the model could be extended to non-native production as there were cases where categorization patterns with perceptual overlap resulted in both discrimination difficulty and less stable non-native vowel production.

In sum, the results from this small-scale study indicate that there is indeed a relationship between perception and production at the initial state of learning. The findings of the present study are in line with the L2LP model which posits that L2 production at the initial state is largely similar to the learner's own native vowel productions. Furthermore, AusE participants also had difficulty producing separate categories for BP contrasts that were perceptually difficult to discriminate. However, our production findings are only based on measurements of acoustic similarity and it would also be beneficial to have native BP speakers rate these tokens to determine the degree of foreign accent in these vowel productions. Finally, the L2LP suggests that L2 production will improve as listeners' perception improves and future studies should test the L2LP claims for L2 development as the participants in the present study were all naïve to BP and therefore representative of the initial state of learning.

5. Acknowledgements

This work was supported by an Australian Postgraduate Award, the MARCS Research Training Scheme and the Australian Research Council Centre of Excellence for the Dynamics of Language (CE140100041).

6. References

[1] Rallo Fabra, L. and Romero, J. "Native Catalan learners' perception and production of English vowels," J. Phon., 40(3), 491–508, 2012.

[2] Piske, T., MacKay, I.R.A. and Flege, J.E. "Factors affecting degree of foreign accent in an L2 : a review," J. Phon., 29(2), 191–215, 2001.

[3] Flege, J.E. "Second language speech learning: Theory, findings, and problems," in Speech perception and linguistic experience: Issues in cross-language research, W. Strange, Ed. Timonium, MD: York Press, 233-276, 1995.

[4] C. T. Best, "A direct realist perspective on cross-language speech perception.," in Speech perception and linguistic experience: Issues in cross-language research, W. Strange, Ed. Timonium, MD: York Press,171-204, 1995.

[5] Escudero, P. "Linguistic Perception and Second Language Acquisition," PhD Dissertation, Utrecht University, 2005.

[6] van Leussen, J.-W. and Escudero, "Learning to perceive and recognize a second language: the L2LP model revised," Front. Psychol., 6, 1–12, 2015.

[7] Levy E.S. and Law F.F. "Production of French vowels by American-English learners of French: language experience, consonantal context, and the perception-production relationship.," J. Acoust. Soc. Am., 128(3), 1290–305, 2010.

[8] Elvin, J., Escudero, P., Williams, D., Shaw, J.A. and Best, C.T. "The role of acoustic similarity, non-native categorisation and individual differences in predicting non-native perception: Brazilian Portuguese vowels by English vs. Spanish listeners," under review

[9] Escudero, P., Boersma, P., Rauber, A. and Bion, R. "A cross-dialect acoustic description of vowels: Brazilian and European Portuguese.," J. Acoust. Soc. Am., 126(3), 1379–93, 2009.

[10] Kisler, T., Schiel, F. and Sloetjes, H. "Signal processing via webservices: the use case WebMAUS," in Proc. of Digital Humanities, 30-34, 2012.

Lebanese Arabic listeners find Australian English vowels easy to discriminate

Ronda Aboultaif[1], *Jaydene Elvin*[1, 2], *Daniel Williams*[1, 2, 3] *& Paola Escudero*[1, 2]

[1]The MARCS Institute for Brain, Behaviour and Development,
Western Sydney University, Australia
[2]The ARC Centre of Excellence for the Dynamics of Language, Canberra, Australia
[3]Linguistics Department, Area of Excellence–Cognitive Sciences, University of Potsdam, Germany
{r.aboultaif, j.elvin, paola.escudero}@westernsydney.edu.au,
daniel.phillip.williams@gmail.com

Abstract

The present study investigated the role of acoustic similarity in predicting bilingual Lebanese Arabic-English (LA) listeners discrimination of Australian English (AusE) vowels. The findings are in line with predictions based on acoustic similarity. In particular, LA listeners use duration as a cue to facilitate vowel discrimination which seems to yield few problems with AusE contrasts, regardless of their L2 proficiency. Furthermore, for LA listeners, discrimination difficulty is only apparent for two contrasts where the vowels do not align with the LA counterparts and when partial acoustic overlapping is identified.

Index Terms: L2 speech perception, vowel discrimination, acoustic similarity

1. Introduction

Second language (L2) learners often struggle with the acquisition of a new sound system. Vowels are particularly difficult to perceive due to the influence of their own native (L1) vowel inventory. Theoretical models such the Perceptual Assimilation Model (PAM) [1], its extension PAM-L2 [2] and the Second Language Linguistic Perception Model (L2LP) [3] posit that L2 learners and naïve listeners perception of non-native or L2 sounds is filtered by their L1, which can lead to difficulties in acquiring the L2.

The L2LP theoretical framework considers three types of learning scenarios that a learner may face in their acquisition of L2 vowels. These scenarios can be identified by investigating the acoustic similarity between the L1 and target L2 [3, 4]. The first scenario, known as the new scenario in L2LP [3, 4] and single category assimilation in PAM [1] occurs when two L2 sounds are mapped to one single native contrast. This generally occurs in learners whose L1inventory is smaller than that of the target language and results in poor discrimination. In contrast the similar scenario in L2LP [5] and two-category assimilation in PAM [1], occurs when the two sounds in the L2 contrast are mapped to two different L1 categories and is generally easy to discriminate. The third scenario, known as the subset scenario in L2LP [6] and uncategorized assimilation in PAM [1], occurs when two non-native vowels in a binary contrast are perceived as belonging to more than two native vowel categories, which is common when the non-native vowel inventory is smaller than that of the native vowel inventory. In this scenario, the L2LP model predicts that discrimination may be problematic when this scenario leads to a "subset" problem where a learner needs to realize on the basis of positive evidence alone that some features or categories in their native language do not exist in the target language and may find it difficult not to perceive an "extra" L1 category [6]. Furthermore, cases of the subset scenario may be particularly difficult to discriminate when both vowels in the L2 contrast are mapped to the same multiple L1 categories, resulting in a perceptual overlap.

While PAM and PAM-L2 [1, 2] typically rely on perceptual assimilation results to predict discrimination accuracy, the L2LP model explicitly states that non-native vowel discrimination can be predicted by a detailed acoustic comparison between the native and target language [3, 4]. In the L2LP framework, a listener's perception and production of non-native or L2 sounds at the initial state of learning should match the acoustic properties of the sounds in their native language [3, 4]. Recent studies have shown acoustic similarity to successfully predict non-native and L2 vowel perception. For example, [7] showed that the cross-language comparison of acoustic properties to successfully predict discrimination accuracy for naïve Iberian Spanish and Australian English listeners of the Brazilian Portuguese.

The aim of the present study is to investigate bilingual Lebanese Arabic-English (LA) listeners' discrimination of Australian English (AusE) vowels. Unlike previous studies investigating native Arabic speakers, e.g. [8], the participants in this study were of a homogeneous background, namely LA, with AusE as their L2. The LA variety that we use to make our acoustic predictions in the present study is the colloquial Arabic spoken in Israel, in the Galilee region [9]. The Israel Arabic belongs to the Levantine dialect group as does LA and have the same five vowel pairs [10]. We use the vowels reported in [9] because these Arabic varieties are very similar. The vowels reported in [9] provide the best approximation to LA vowel acoustics including formant values, as LA is a dialect which is very much understudied and the descriptions which are currently available are either outdated or incomplete, e.g. [11]. The LA vowel inventory consists of ten monophthongs /i, iː, a, aː, u, uː, e, eː, o oː/ [9] and the AusE vowel inventory consists of twelve monophthongs, namely, /iː, ɪ, e, eː, ɜː, ɐ, ɐː, æ, oː, ɔ, ʊ, ʉː/ [12]. Both languages employ phonemic length to distinguish short and long vowels (e.g. /a/-/aː/ in LA and /ɐ/-/ɐː/ in AusE) [9, 12]. Importantly, LA has five vowel contrasts that differ by length only, while many AusE vowels differ by both vowel quality and phonemic length. Thus, this study will also investigate whether LA listeners' use duration as a cue to discriminate AusE vowel contrasts, as studies [e.g., 5] have shown L2 learners to use durational differences in L2 vowel categorization.

Given that studies [e.g, 7] have shown that acoustic similarity between the native and target language successfully predicts L2 and non-native vowel discrimination, we apply a similar method of prediction to the present study. Figure 1 shows the F1 and F2 values for LA and AusE monophthongs.

In this preliminary study, predictions based on acoustic similarity for the six target AusE contrasts /iː/-/ɪ/, /e/-/eː/, /ɜː/-/æ/, /ɐ/-/ɐː/, /oː/-/ɔ/ and /ʊ/-/ʉː/ were made by visually comparing the AusE and LA vowels in the acoustic space.

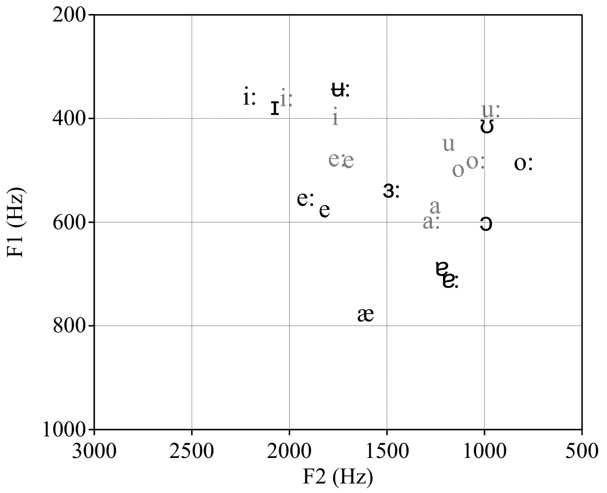


Figure 1: *Average F1 & F2 acoustic values for vowels produced by native AusE males in black [13] and LA males in grey [9].*

As shown in Figure 1, the AusE contrasts /ɐ/-/ɐː/and /e/-/eː/ seem to be acoustically similar to the LA vowel contrasts /aː/-/a/ and /e/-/eː/, which also differ by length only and this should facilitate the discrimination of these vowel contrasts. For the AusE /æ/-/ɜː/ contrast, although AusE /æ/ seems acoustically distant from most LA vowels, it is potentially acoustically similar to the LA vowels /a/ and /aː/. On the other hand, AusE /ɜː/ appears to be in close acoustic proximity to LA /aː/, /aː/, /eː/ and /e/. The fact that there is acoustic overlap with both vowels in the AusE /æ/-/ɜː/ being acoustically similar to LA /a/ and /aː/, discrimination accuracy may be lower for this vowel contrast. However, the lip rounding of AusE /ɜː/ and if the participants use duration as a cue, may result in fewer discrimination difficulties for this contrast.

Duration may be used as a cue to facilitate the discrimination of the AusE /iː/-/ɪ/, but both vowels are acoustically closer to LA /iː/ than /i/. This may result in discrimination difficulties for LA listeners if they perceive both vowels in the contrast as LA /iː/. In the case of AusE /oː/-/ɔ/, we expect fewer discrimination difficulties as AusE /oː/ appears acoustically close to a similar LA long vowel /oː/ and is also acoustically similar to the LA long vowel /uː/, while AusE /ɔ/ is in acoustic proximity to LA short vowels /a/ and /o/.

LA listeners may not find AusE /ʊ/-/ʉː/ difficult to discriminate because a similar short-long vowel contrast exists in LA. Furthermore, we observe that AusE /ʊ/ is acoustically close to LA /u/, /uː/, /o/, /oː/, while AusE /ʉː/ is much more fronted. Although AusE /ʉː/ appears acoustically closer to LA /iː/ in terms of F1 and F2 values, it is unlikely that this vowel will be perceived as LA /iː/, as it is a rounded vowel.

In sum, LA listeners, whose native vowel inventory includes phonemic length contrasts, should use duration as a cue that facilitates L2 vowel discrimination. However, discrimination may be more difficult for those contrasts where there is a vowel quality difference between AusE and LA vowels and when both AusE vowels are mapped to one or more of the same LA categories. A summary of the possible learning scenarios for LA learners of AusE are provided in Figure 2.

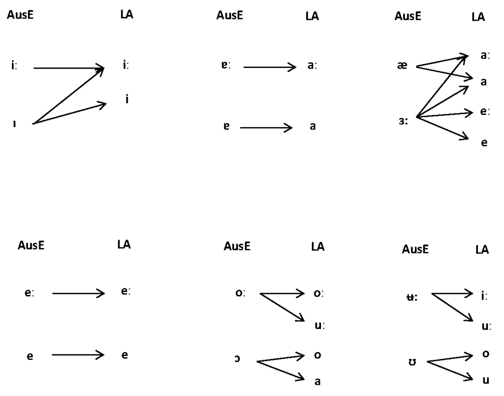


Figure 2: *Summary of possible learning scenarios for native LA listeners' discrimination of the six AusE contrasts, according to the L2LP model and based on visual acoustic comparison.*

2. Method

2.1 Participants

Listeners were 15 native LA-English bilinguals and 15 AusE monolinguals, aged between 18 and 45 (mean age 35.9 for LA, 24.8 for AusE). The LA listeners were born in Lebanon; migrated to Australia within the last 18 years and currently live in the suburbs of Western Sydney. LA participants spoke AusE as a L2 and reported either low-intermediate or advanced proficiency in English in a language background questionnaire administered prior to testing. The AusE participants were born in Australia, reported little to very basic knowledge of any foreign language and were included as our control group.

2.2 Stimuli and Procedure

Stimuli for the XAB task were 120 AusE natural isolated vowel tokens produced by 5 male and 5 female monolingual speakers of AusE from Western Sydney selected from the [14] corpus. There were 10 tokens for each of the 12 AusE vowels /iː, ɪ, e, eː, ɜ, ɐ, ɐː, æ oː, ɔ, ʊ, ʉ/ (12 AusE x 10 repetitions) extracted from nonce words that all rhymed with real words, produced in the /fVf/ context [13]. We also extracted 12 natural vowel tokens (representing each of the 12 AusE vowels) produced by 1 male and 1 female AusE monolingual speaker to use as the A and B stimuli.

As in [7], participants were presented with an auditory discrimination task presented in an XAB format. Participants completed six blocks, each containing 40 trials. These six blocks represented the six AusE long vs short vowel contrasts, /æ/-/ɜː/, /e/-/eː/, /iː/-/ɪ/, /ʊ/-/ʉː/, /ɐ/-/ɐː/ and /oː/-/ɔ/. In each trial, listeners heard three vowel sounds, one after the other and were asked to decide whether the first (X) sounded more like the second (A) or third (B), by pressing the corresponding button on the keyboard. The order of the A and B stimuli were counterbalanced and to ensure language-specific phonological processing, the inter-stimulus interval was set to 1.2 seconds.

Participants were tested at Western Sydney University and were first presented with a practice block to ensure that they understood the task. Instructions were provided in English and the entire testing session took approximately 45 minutes to complete.

3. Results

Figure 3 shows the accuracy scores for the LA and AusE listeners across the six AusE vowel contrasts. These results indicate that both listener groups found AusE /ʊ/-/ʉː/ the easiest to discriminate, while AusE listeners found the AusE /e/-/eː/ contrast the most difficult contrast to discriminate, whereas it seems that LA listeners found both AusE /æ/-/ɜː/ and /e/-/eː/ the most difficult. To determine whether accuracy scores differed significantly across the two groups and across the different contrasts, we ran a repeated-measures ANOVA with language as a between-subjects factor and vowel contrast as a within-subjects factor. The results indicated a main effect of contrast [F (5,140) = 44.995, p = < 0.001 $\eta_p{}^2$ = 0.616] and a contrast*language interaction [F (5,140) = 7.727, p = < 0.001, $\eta_p{}^2$= 0.616], but no main effect of language (p = 0.949).

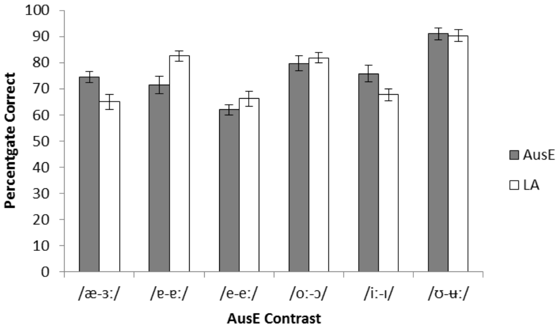


Figure 3. *Discrimination accuracy for six AusE vowel contrasts by native AusE and native LA bilingual English listeners.*

To test the predictions for each contrast stated in the introduction, we ran planned comparisons using independent-samples t-tests comparing AusE and LA listeners' accuracy. We found significant differences for the /æ/-/ɜː/, /ɐ/-/ɐː/ and /iː/-/ɪ/ vowel contrasts. In line with our predictions, LA listeners had significantly lower discrimination accuracy than AusE listeners for /æ/-/ɜː/ (t(28) = -2.661, p = 0.013) and /iː/-/ɪ/ (t(28) = -2.044, p = 0.05), while they had higher accuracy for /ɐ/-/ɐː/ (t(28) = 2.841, p = 0.008).

The results above may be influenced by the difference in L2 proficiency within the LA group. We therefore split LA listeners into two groups according to their self-evaluation of English proficiency (e.g. low, intermediate or advanced). Participants were included in the low group (LA_1) if they indicated that their English proficiency level was at an intermediate level or lower and participants included in the high group (LA_2) were those who indicated that their English proficiency was at an advanced level or higher. The LA_1 group (low proficiency) included 8 listeners and the LA_2 group (high proficiency) included 7 listeners. Figure 4 shows the two LA groups differences in discrimination accuracy of the six AusE vowel contrasts.

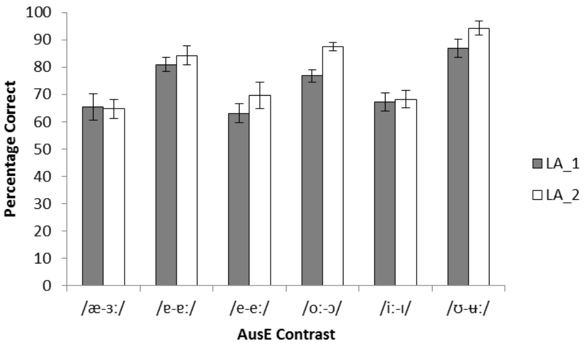


Figure 4. *Discrimination accuracy for six AusE vowel contrasts by 8 LA intermediate L2 English listeners (LA_1) and 7 LA advanced (LA_1) L2 English listeners.*

Figure 4 shows that the LA_2 group had higher overall discrimination accuracy and that both groups had high accuracy scores for AusE /ʊ/-/ʉː/. The results suggest that the LA_1 group found the AusE /e/-/eː/ most difficult to discriminate, whereas the LA_2 group found AusE/æ/-/ɜː/ the most difficult. To determine whether or not the results observed in Figure 4 are statistically significant, we ran a repeated-measures ANOVA with proficiency group as a between-subjects factor and contrast as a within-subjects factor. While the results indicated a main effect of contrast, [F (5,65) = 26.604, p = <0.01, η2p = 0.672], no effect of proficiency (p = 0.329) or interaction between contrast*proficiency (p = 0.404) was found. Although it appears that the LA_2 group performed better overall, this finding did not reach significance, indicating that AusE proficiency does not play a role in LA listeners' discrimination accuracy.

4. Discussion

The present study investigated LA listeners' discrimination of six AusE vowel contrasts. Following the L2LP framework, we compared the acoustic similarity between the native and target language to predict L2 discrimination difficulty.

The results indicate that LA listeners do indeed rely on duration as a cue to facilitate their discrimination between AusE short and long vowels. In fact, LA listeners performed better than native AusE listeners on the /ɐ/-/ɐː/ contrast which differs by length only. Interestingly, LA listeners' accuracy was lower than AusE listeners on the vowel contrasts that do not perfectly align to the vowels in their own native length contrasts and in which a partial acoustic overlap was identified (e.g., /æ/-/ɜː/ and /iː/-/ɪ/). These findings suggest that LA listeners' performance is equal to and in some cases better than native speakers when the contrast is similar to their own native contrasts, with no acoustic overlapping. However, duration may not be a sufficient to facilitate the discrimination of L2 vowel contrasts that contain vowel quality differences as well as an acoustic overlap. In addition, we found that the LA listeners' proficiency in AusE did not facilitate their vowel discrimination, suggesting that their perception of AusE is still strongly influenced by their L1.

The findings from the present study are in line with the L2LP model's claim that L2 perceptual difficulty can be predicted by the acoustic similarity between the native and target language. That is, LA listeners did indeed find AusE /ʊ/-/ʉː/ easy to discriminate and AusE /æ/-/ɜː/ and /iː/-/ɪ/ the most difficult to discriminate. Although our findings were not consistent with our predictions for AusE /e/-/eː/, native AusE listeners also found this contrast difficult to discriminate and it may be that this difficulty is a result of the fVf context in which these two vowels were extracted from, namely *fairf* and *fef* where the length difference between the vowels may be subtle. That is, the vowels may be better identified in a "consonantal syllabic context" rather than as an isolated vowel, where acoustic information is not completely captured influencing identification accuracy [14].

In sum, our findings suggest that LA learners of AusE use their native vowel length contrasts to facilitate L2 vowel discrimination. This use of vowel duration differences seem to yield very few problems with AusE contrasts in LA listeners who, in fact, outperform AusE listeners in one contrast, regardless of their proficiency with AusE. For LA listeners, discrimination difficulty is only apparent for two of the six AusE contrasts where the vowels do not align with the LA counterparts and when partial acoustic overlapping is identified.

The present results can only be considered preliminary for the following reasons. First, our acoustic predictions were limited because we used published LA acoustic data, which may not entirely reflect the LA dialect of the listeners tested in this study and our predictions were based on visual inspection of the vowel plot. Further investigation is required that provides a more accurate measure of acoustic similarity that also considers vowel duration and the relative importance of formant frequencies and vowel duration in AusE and LA. Second, the difference in mean age between AusE and LA listeners may have influenced the results and should be controlled for with a new sample of AusE participants. Finally, the effect of proficiency was examined with groups of less than 10 participants each. A larger sample per proficiency group may yield an influence of L2 proficiency on the accuracy with which LA listeners discriminate AusE vowel contrasts.

5. Acknowledgements

This research was supported by the Australian Research Council Centre of Excellence for the Dynamics of Language (CEI40100041).

6. References

[1] Best, C., "A direct realist view of cross-language speech perception. Speech perception and linguistic experience: Issues in cross-language research", pp.171-204). http://www.haskins.yale.edu/Reprints/HL0996.pdf, 1995.

[2] Best, C. T., and Tyler, M. D., "Nonnative and second-language speech perception: Commonalities and complementarities", in O.–S. Bohn and M. Munro [Eds], Language experience in Second language speech learning: in Honor of James Emil Flege, pp.13-34, John Benjamins, Amsterdam, 2007.

[3] Escudero, P., Linguistic perception and second language acquisition: explaining the attainment of optimal phonological categorization. PhD thesis, Utrecht University, he Netherlands, 2005.

[4] Escudero, P., "The linguistic perception of similar L2 sounds", in P. Boersma and S.Hamann [Eds], Phonology in Perception, pp.152–190, Mouton de Gruyter, Germany, 2009.

[5] Escudero, P., and Boersma, P., "Bridging the gap betweenL2 speech perception research and phonological theory," Studies in Second Language Acquisition, 26(5), pp.51–585. doi: 10.1017/S027226310404002, 2004.

[6] Escudero, P., and Boersma, P. "The subset problem in L2 perceptual development: Multiple-category assimilation by Dutch learners of Spanish, in Proceedings of the 26th annual Boston University Conference on Language Development, edited by B. Skarabela, SFish, and A. Do (Cascadilla Somerville), pp. 208-219, 2002.

[7] Elvin, J., Escudero, P., and Vasiliev, P., "Spanish is better than English for discriminating Portuguese vowels: Acoustic similarity versus vowel inventory," Frontiers in Psychology, Language Sciences, 5(1188), 2014.

[8] Shafiro, V., Levy, E. S., Khamis-Dakwar, R., and Kharkhurin, A., "Perceptual confusions of American-English vowels and consonants by native Arabic bilinguals," Language and Speech, 56(2), pp.145-161. doi: 10.1177/0023830912442925, 2013.

[9] Amir, N., Amir, O., and Rosenhouse, J., "Colloquial Arabic vowels in Israel: A comparative acoustic study of two dialects," The Journal of the Acoustical Society of America, 136(4), pp.1895-1907, 2014.

[10] Amir, N., Amir, O., and Rosenhouse, J., "Vowel systems of quantity languages compared: Arabic dialects and other languages," Acoustical Society of America, 2, pp.1-10. doi:10.1121/1.4880205, 2014

[11] Abu-Haidar, F. "A Study of the spoken Arabic of Baskinta". EJ Brill, Leiden-London, 1979.

[12] Cox, F., and Palethorpe, S., "Australian English," Journal of the International Phonetic Association, 27(03), pp.341-350. doi:10.1017/S0025100307003192, 2007.

[13] Elvin, J., Williams, D., and Escudero, P., "Dynamic acoustic properties of monophthongs and diphthongs in Western Sydney Australian English," Journal of the Acoustical Society of America-EL, 140(1), pp. 576-581. 2016.

[14] Strange, W., Edman, T, R., Jenkins, J, J., "Acoustic and phonological factors in vowel identification," Journal of Experimental Psychology: Human Perception and Performance, 5(4), pp.643-656. 1979.

The Effect of Sampling Procedures on the Performance of Likelihood Ratio Based Forensic Voice Comparison: F0 Distributional Parameters

Shunichi Ishihara

Department of Linguistics, the Australian National University

shunichi.ishihara@anu.edu.au

Abstract

In this study, various sampling procedures were tested for F0 distributional parameters in order to see how the performance of the forensic voice comparison system is influenced. This was done by changing the width of analysis window and the degree of its shifting, resulting in different sets of feature vectors. The results show that the discriminability is fairly comparable across the different sampling procedures, whereas there is a large difference in calibration loss between the procedures if the analysis windows are not overlapped. It is reported that the use of overlapped analysis windows contributes to the stability in calibration loss.

Index Terms: forensic voice comparison, long-term F0 distribution, sampling procedures

1. Introduction

The usefulness and efficacy of the features based on long-term F0 distributional patterns (e.g. the mean, sd, skewness, kurtosis of the distribution) have been reported and demonstrated for forensic voice comparison (FVC) [1]. However, since the long-term F0 distribution is usually obtained from the entire speech sample available for the caseworker, only one set of feature vectors is consequently available for modelling. It goes without saying that the more sets, the better for accurately estimating the variance of a speaker [2]. Thus, a question naturally arises: isn't it better to obtain multiple numbers of feature vectors for each speaker by breaking up the recording into some chunks? Needless to say, there is a trade-off here between the amount of data for building the distributional model of each chunk and the number of the set of feature vectors. This study attempts to answer the above question, with the Multivariate Kernel Density (MVKD) likelihood ratio formula [3], which is one of the common procedures for FVC.

2. Likelihood ratio

The current study is a likelihood ratio (LR) based FVC study. For FVC, as expressed in Equation (1), LR is the probability of observing the difference (referred to as the evidence, *E*) between the offender's and the suspect's speech samples if they had come from the same speaker (H_p = the prosecution hypothesis, is true) relative to the probability of observing the same evidence (E) if they had been produced by different speakers (H_d = defence hypothesis, is true) [4].

$$LR=\frac{p(E|H_p)}{p(E|H_d)} \tag{1}$$

The LR expresses the relative strength of the given evidence regarding the competing hypotheses (H_p vs. H_d). It is a common practice to present LRs in the logarithmic scale (base 10), in which case the neutral point is the $\log_{10}$LR = 0.

3. Database and F0 extraction

The monologues stored in the Corpus of Spontaneous Japanese (CSJ) [5] are used in this study. The criteria for selecting speakers from the CSJ, the pre-processing (e.g. downsampling) of the selected speech samples and so on, are explained in detail in [1]. The selection criteria resulted in the selection of 201 speakers (201 speakers * 2 non-contemporaneous sessions = 402 speech samples), and they were divided into three mutually exclusive databases of test, background and development (each of which consists of 67 speakers).

The stream of speech in a recording is pre-annotated and chunked into the unit of *utterance* in the CSJ. Utterances are separated by silences with durations of 0.4 sec. or longer. The CSJ also annotates non-speech noise, and the sections marked with a noise tag were excluded from the F0 extraction. The ESPS routine of the Snack Sound Toolkit (http://www.speech.kth.se/snack/) was used to extract F0 at every 0.005 second from the utterances for each recording. The distributional pattern of the extracted F0 values was parameterised by calculating the following six features: the mean, standard deviation, skew, kurtosis, modal F0 and the density of the modal F0. The KernSmooth library of the R statistical package was used for the modal F0 and its density.

4. Sample sizes and sampling procedures

As can be seen in Table 1, eight different sample sizes are used in this study. They are given in terms of the numbers of F0 samples and their equivalent durations.

Table 1: *F0 sample sizes. Durations in sec.*

Numbers	Durations
2000	10
4000	20
6000	30
8000	40
12000	60
16000	80
20000	100
24000	120

For example, 2000 means that the first 2000 F0 values of a given recording are used to build a long-term F0 distribution, from which the six features are extracted. Since the F0 values were calculated at every 0.005 sec., the duration of the sample is equivalent to 10 sec. (= 2000 * 0.005). In order to avoid unnecessary confusion, the different samples sizes are referred to with their durations in sec.

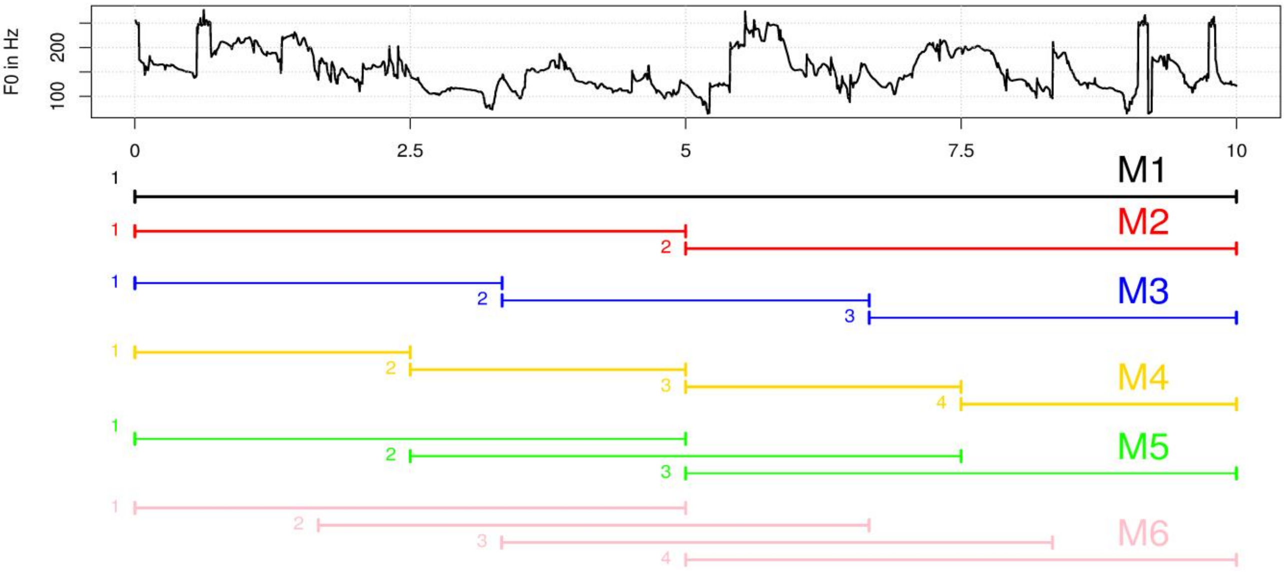


Figure 1: *Six sampling procedures (M1-M6) are schematically presented with sequentially plotted F0 values (x-axis = sec.).*

Six different sampling procedures are tried in this study. They (M1-M6) are schematically given in Figure 1. In the top half of Figure 1, the first 2000 F0 values extracted from an example recording are sequentially plotted (10 sec. in total). As can be seen in Figure 1, the six different sampling procedures have different analysis window sizes and different degrees of shifting (or overlaps), resulting in different amounts of feature vector sets. The F0 values which fall within each analysis window are pooled together in order to build the long-term F0 distribution. In M1-M4, for example, i) the size of each analysis window is different and ii) there is no overlap between the adjacent windows. In M1 (the benchmark case), the maximum size of analysis window is used; that is, only one set of feature vector. In M2, M3 and M4, the 50%, 33.3% and 25% of the maximum window size are used, resulting in two, three and four sets of feature vectors, respectively. In M5 and M6, half of the maximum window size is used with the analysis window being shifted by 50% and 33.3%; three and four sets of feature vectors, respectively.

5. Estimation of strength of evidence and performance assessment

The Multivariate Kernel Density (MVKD) formula [3] was used with a logistic-regression calibration [6] to estimate the likelihood ratios (LRs). In the MVKD formula, the covariance matrices (D_l, l = 1,2) of the offender and the suspect samples are assumed to be constant, and they are estimated from the pooled within-speaker covariance matrix (U) of the background database, being scaled by the number of samples (= feature vectors) (n) ($n_l{}^{-1}U,\ l = 1{,}2$). That is, only the suspect and offender means are used in the calculation of the LRs in the MVKD formula. Thus, when the difference between the feature vectors is compared by means of a *Mahalanobis* distance, the number of feature (n) vectors plays an important role.

A logistic-regression calibration is applied to the outcomes (customarily called *scores*) of the MVLR formula. The logistic-regression weight is obtained from the development database. For the different sample sizes, the performances of the six different sampling procedures given in Figure 1 are assessed by means of the log-likelihood-ratio-cost (C_{llr}), including C_{llr}_*min* and C_{llr}_*cal*. The C_{llr} measures the overall accuracy of an FVC system. The C_{llr}_*min* and C_{llr}_*cal* specifically examine the discrimination and calibration performances of the system, respectively [6]. The FoCal toolkit (https://sites.google.com/site/nikobrummer/focal) was used for both the calibration and C_{llr}.

6. Results and discussion

The C_{llr}, C_{llr}_*min* and C_{llr}_*cal* values are given in Figure 2 for each of the sampling procedures (M1-M6). The left-hand side of Figure 2 (Panels a, b and c) are for M1-M4, and the right-hand side (Panels d, e and f) are for M1, M5 and M6. M1 is the benchmark experiment. The different sizes of analysis window were used in M1, M2, M3 and M4 with 100% shift (= no overlapping between the windows). Half of the maximum window size was used in both M5 and M6 with different degrees of shifting (50% and 33.3%, respectively).

Judging from Figure 2a, in which the C_{llr} values of M1-M4 are given, it is not that one procedure consistently performs better than the others; yet M2 (red) performed better than M1 (black) for all sample sizes, except the sample size of 80 sec. The performances of M1-M4 are relatively comparable up to the sample size of 40 sec., after which they largely fluctuate even within the same sampling procedures, resulting in that the best- or worst-performing sampling procedure is not consistent; for example, M1 achieved the best C_{llr} (= 0.613) for the sample size of 80 sec. while M1 performed worst (C_{llr} = 0.775) with the sample size of 100 sec. As for the sample sizes of 100 and 120 sec., M2-M4 (multiple sets of feature vectors) consistently performed better than M1 (one set of feature vector). An advantage of having multiple sets of feature vectors may start emerging when the sample size is relatively large (100 sec. or longer); perhaps each chunk is large enough to accurately estimate the distributional pattern. This point needs to be confirmed with a larger amount of data.

The observation of Figure 2bc, in which C_{llr}_*min* and C_{llr}_*cal* values are plotted, respectively, as a function of the sample size, tells that i) the discriminability of the system (Figure 2b) is fairly comparable across M1-M4, ii) the discriminability of the system (Figure 2b) generally improves

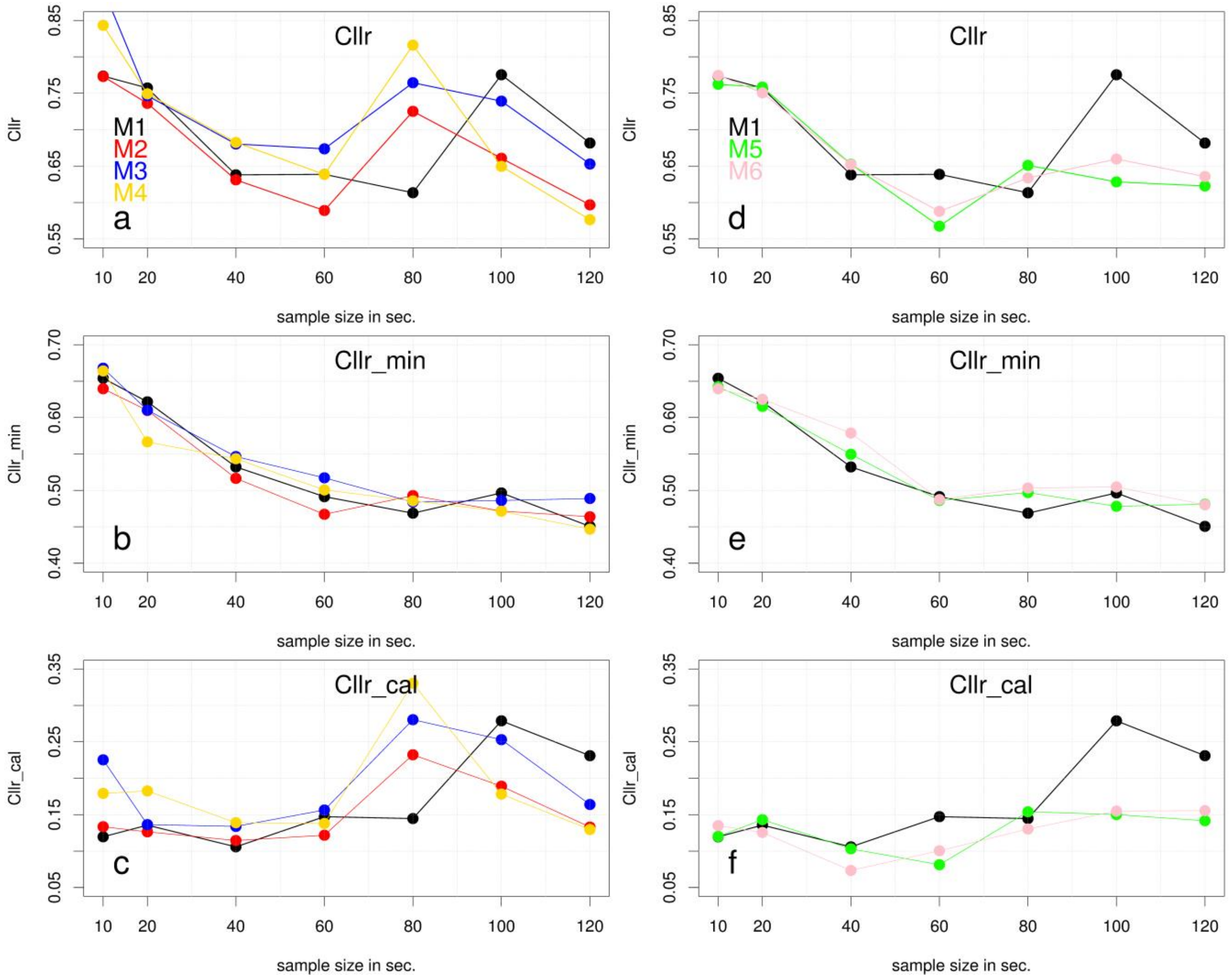


Figure 2: *The C_{llr} (Panels a and d), C_{llr}_min (b and e) and C_{llr}_cal (c and f) values of the different sampling procedures (M1-M6) are plotted against the different sample sizes (10, 20, 40, 60, 80, 100 and 120 sec.). Panels a, b and c are for M1-M4, and Panels d, e and f are for M1, M5 and M6. The M1 is the benchmark experiment.*

as the amount of sample size increases until the sample size of 60 sec., after which the performance starts converging; this is something we naturally expect and iii) the calibration loss (Figure 2c) is not stable (the sample size of 80 sec. or longer).

It is interesting to know that the large fluctuations in performance in terms of C_{llr}, which is demonstrated in Figure 2a, in particular when the sample size is 60 sec. or longer, are due to large fluctuations in calibration performance. Although the reason behind this is not clear at this stage, it may be due to an inherited nature of the MVKD formula [7].

The results of M5 and M6 are plotted in Figure 2def together with the results of M1. M5 and M6 are different from M1-M4 in that the analysis window shifted in the way that adjacent windows overlap; the degree of overlap is 50% and 66.6%, respectively. Needless to say, overlapped analysis windows have a smoothing effect in sampling.

It is clear from Figure 2d that the performances of M5 and M6 are very similar in terms of C_{llr}, and also that the performance improves more or less in an expected manner as the sample size increases. This observation of M5 and M6 is quite different from that of M1-M4, which is not as stable. Furthermore, like the observation made for M1-M4 in Figure 2a, M5 and M6 performed better than M1 when the sample size is large (e.g. 100 and 120 sec.).

As for the discriminability potential of M5 and M6 (refer to Figure 2e), it shows the same trend as that of M1-M4. Unlike M1-M4, the calibration loss of M5 and M6 is relatively stable, as shown in Figure 2f, in that the C_{llr}*_cal* values sway around the narrow range between C_{llr}*_cal* = 0.05 and 0.15. It is clear that overlapped analysis windows contribute to the stability of the system performance.

Figure 3 contains the Tippett plots of M1, M2, M5 and M6 for the sample size of 120 sec. In terms of C_{llr} values, M2 (0.576), M5 (0.622) and M6 (0.635) are fairly similar in performance, but they are better in performance than M1 (0.681). It is evident from Figure 3 that some large counter-factual same speaker LRs, which are indicated by the circle in Figure 3a, largely influence the C_{llr} value; otherwise the magnitude of the derived LRs, including both consistent-with-fact and contrary-to-fact LRs, are fairly similar across the different sampling procedures. This observation is generally true for the other cases.

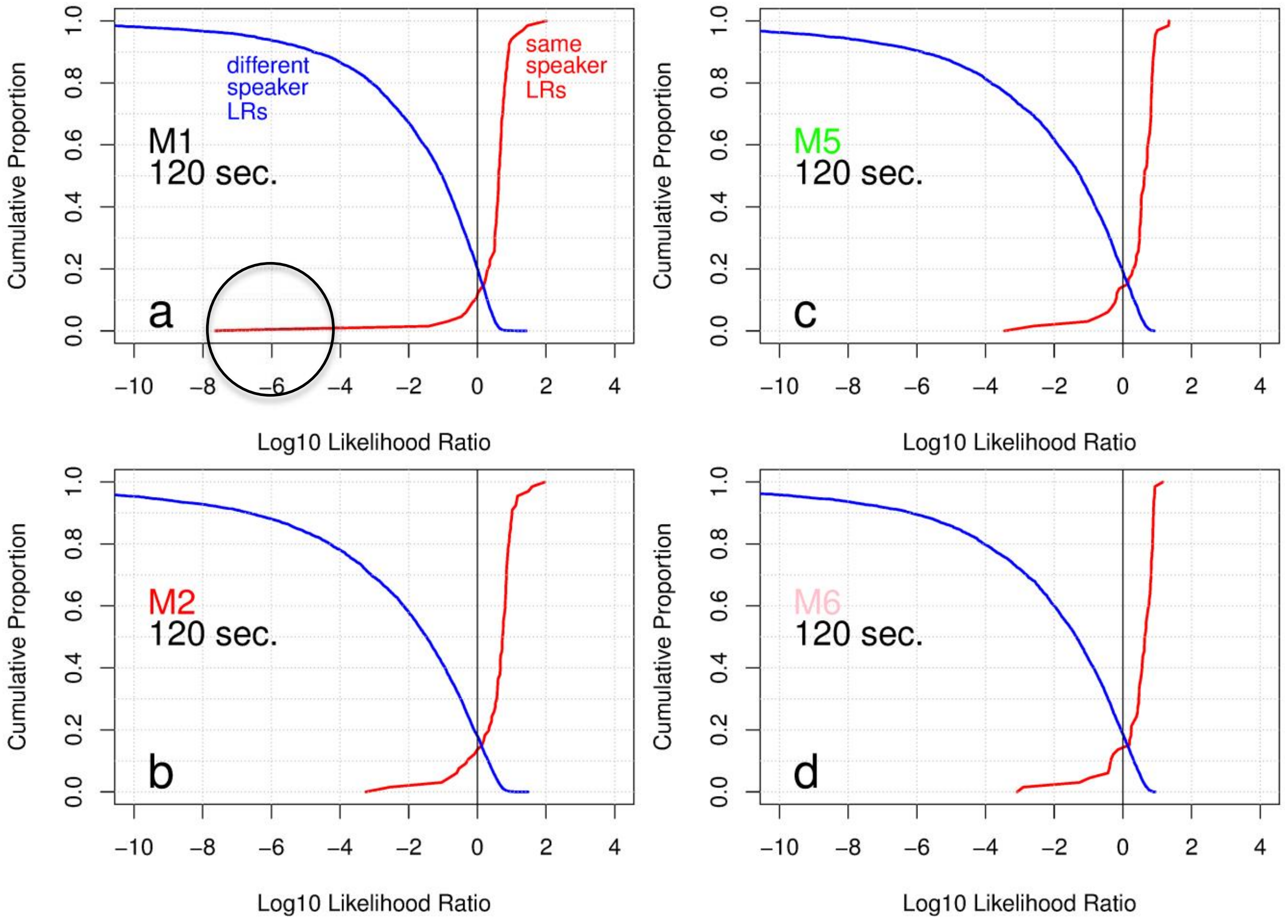


Figure 3: Tippett plots of M1, M2, M5 and M6 for the sample size of 120 sec. Red and blue curves = same speaker and different speaker LRs, respectively. The circle indicates large counter-factual same speaker LRs.

7. Conclusions

As far as the results of the current study are concerned, the positive effect of the multiple sets of feature vectors was not consistently observed, nevertheless, it was also pointed out i) that the multiple sets of feature vectors appear to perform better than the benchmark (M1) when the sample size is large (e.g. 100 sec. or longer) and ii) that M2 (two sets of feature vector) consistently performed better than M1, except when the sample size is 80 sec. An interpretation of the result of the present study is that the caseworker should try some different sampling procedures to see how the derived LR fluctuates.

It is theoretically interesting to know that the discriminability was relatively comparable across the different sample sizes (M1-M6), whereas the calibration loss largely fluctuated depending on the sample size if the analysis windows are not overlapped (M1-M4). The effect of the overlapped analysis windows is evident in that the calibration loss becomes fairly stable.

It is also interesting to see how the performance of the FVC system will improve when the F0-based LRs are fused with the LRs obtained from more powerful filter-based features (e.g. formant frequencies).

8. Acknowledgements

I would like to thank the two anonymous reviewers for their valuable comments.

9. 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] S. Ishihara, "The Effect of the Within-speaker Sample Size on the Performance of Likelihood Ratio Based Forensic Voice Comparison: Monte Carlo Simulations," in *Australasian Language Technology Association Workshop 2013*, Brisbane, Australia, 2013, pp. 25-33.

[3] 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.

[4] B. Robertson and G. A. Vignaux, *Interpreting Evidence: Evaluating Forensic Science in the Courtroom*. Chichester: Wiley, 1995.

[5] K. Maekawa, H. Koiso, S. Furui, and H. Isahara, "Spontaneous speech corpus of Japanese," in *the 2nd International Conference of Language Resources and Evaluation*, Athens, Greece, 2000, pp. 947-952.

[6] N. Brümmer and J. du Preez, "Application-independent evaluation of speaker detection," *Computer Speech and Language,* vol. 20, pp. 230-275, Apr-Jul 2006.

[7] 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.

Exploring forensic accent recognition using the Y-ACCDIST system

Georgina Brown
University of York, UK
gab514@york.ac.uk

Abstract

Forensic speech scientists may sometimes be faced with the task of extracting information about an unknown speaker in a recording. It is proposed here that accent recognition technology could assist analysts in such cases and we begin to explore the Y-ACCDIST system's potential for this purpose. Research on Y-ACCDIST so far has largely focussed on its ability to distinguish between varieties which are much more similar to one another than previous automatic accent recognition research [1]. The experiments presented here build on this and challenge Y-ACCDIST in other ways relevant to forensic applications: spontaneous speech data and degraded data.

Index Terms: forensic speech technology, accent recognition

1. Introduction

Most of the work by a forensic speech analyst is on *speaker comparison* tasks. This is the task of analysing one speech recording against another to assess whether the speech in both recordings was produced by the same speaker. Speaker recognition technology can assist with this kind of task. On a lesser scale, there is another task a forensic analyst might be faced with called *speaker profiling*. Rather than comparing two samples, this type of task involves analysing a speech sample to extract information about the speaker. One example of the type of real-life case is a ransom telephone call [2]. Useful information might include his or her geographical origin. Currently, the speaker profiling task is done manually with no technology to assist with this kind of task. This paper further investigates the possibility of testing a specific automatic accent recognition system, Y-ACCDIST (the York ACCDIST-based automatic accent recognition system) [1], for forensic applications.

The following addresses the idea of forensic accent recognition by first giving an overview of accent recognition in section 2. It first reviews forensic speaker profiling (from a manual point of view) before then exploring what has already been achieved by speech technologists in automatic accent recognition more generally. Section 3 presents experiments of the Y-ACCDIST system being applied to forensically relevant accent data. This includes a technical description of Y-ACCDIST. Section 4 summarises the paper and puts forward ways in which the research area of automatic accent recognition for forensic applications could be further developed.

2. Accent Recognition

While accent recognition and perception is of course of sociolinguistic interest, this section will initially focus on accent recognition in relation to the forensic context. The second part of this section will review past automatic accent recognition studies, which have not necessarily been motivated by forensic applications, but mostly by automatic speech recognition.

2.1. Forensic Speaker Profiling

Little literature exists on the topic of speaker profiling. As already stated, forensic speech science is largely concerned with the speaker comparison task. This imbalance of research attention is naturally reflective of a forensic analyst's workload. One study which could be loosely tied to the task of speaker profiling is [3]. They assessed the ability of different analysts to classify speakers based on accent, in the context of *Language Analysis for the Determination of Origin* (LADO). LADO is applied to a small proportion of asylum seeker applications which require additional assessment to establish information about the applicant. More specifically, we might want to know if the applicant is from where he or she claims, and an analysis of a recorded interview can assist with this. [3] had recorded stimuli of native Ghanaian English speakers and Nigerian English speakers. Different analysts (with varying degrees of expertise) were asked the question for each speaker: Do you believe this person is speaking Ghanaian English? They compared the performance of academic phoneticians, undergraduate students, native speakers of Ghanaian English and LADO professionals. They found that the native speakers were the highest-performing group at this task with an overall classification rate of 86% correct. It would be of great interest to see how an automatic system would compare in a similar task, and whether human analysts and an automatic system could combine their strengths to obtain an overall higher result. This has not yet been explored. To begin to do this, the following section covers some of the developments of automatic accent recognition systems within speech technology.

2.2. Automatic Accent Recognition

Traditionally, automatic accent recognition research has focussed on building systems which improve the overall performance of automatic speech recognition systems. By identifying the accent of a speaker before attempting to recognise what is being said, we can raise speech recognition rates [4].

A range of approaches have been trialled to conduct automatic accent recognition. Taking inspiration from the work of [5] in Language Identification (LID), [6] applied a variation of a *Phone Recognition followed by Language Modelling* (PRLM) system to the task of distinguishing between different Arabic varieties. By estimating the sequence of phones in the unknown utterances, using a phoneme recogniser, we can then assess this sequence in terms of how likely the sample belongs to each of the varieties in the reference set. [6] suggest that the varieties of Arabic they use are distinguishable by the different phonemic sequences, and this is reflected in the 6.0% Equal Error Rate they achieve.

Other approaches have included more acoustic modelling, rather than phonotactic modelling as above. [7] applied Gaussian Mixture Models (GMMs), combined with Mel -Frequency

Cepstral Coefficients (MFCCs), to the classification problem of distinguishing between four Mandarin Chinese accent varieties. With this approach, they achieved an error rate of 11.7% and 15.5% for females and males respectively.

The approaches discussed so far are text-independent. Text-independent systems come with the practical advantage of not needing an orthographic transcription to accompany the speech sample. As a consequence, this broadens the pool of applications these systems can be used for. However, some applications (like the forensic application) might benefit from a text-dependent option if it can bring the precision that is required. [8] introduced the ACCDIST metric for automatic accent recognition. It is a text-dependent method which calculates intra-speaker distances between vowel sounds found in a speech sample. A clearer idea of how an ACCDIST-based system works is given in the Y-ACCDIST system description further below in 3.1.

[9] compared a number of automatic accent recognition systems on the same corpus. They compared a combination of text-independent and text-dependent systems. These were GMM-based acoustic systems and ACCDIST-based systems. They were all tested on the *Accents of the British Isles* corpus [10]. This corpus contains recordings of speakers from 14 locations spanning the breadth of the British Isles. In a 14-way classification their highest-performing system (which was a text-dependent ACCDIST-based system) achieved 95.18% accuracy. This unsurprisingly outperformed their text-independent GMM-based systems (which achieved 61.13% and 76.11% on the same task). Of course, a text-dependent system outperformed the text-independent systems, but the text-independent systems still seem to perform well on a 14-way classification task.

2.3. Y-ACCDIST

Y-ACCDIST is a text-dependent accent recognition system based on the ACCDIST metric [8]. What separates Y-ACCDIST from previously developed ACCDIST-based systems in [8] and [9] is that Y-ACCDIST is able to process content-mismatched speech. Past ACCDIST-based systems have relied on testing and training speech content to be the same, as the vowel segments they analyse and compare are restricted to specific contexts. In the case of [8], word-specific vowels are analysed and compared, and in the case of [9], triphone-context vowels are used. Y-ACCDIST, on the other hand, collapses phones into phonemic categories to analyse and compare. Details of how this is done are given in section 3.1.

Y-ACCDIST has previously been tested with the forensic application in mind. [11] explored Y-ACCDIST's performance on varieties which were assumed to have a greater degree of similarity between them. Past research on automatic accent recognition (due to the focus on automatic speech recognition) has been concerned with categorising speakers into accent groups with great differences between them. This is of course useful to automatic speech recognition which suffers due to the great variation among speakers. However, for the forensic application, it is of interest to challenge an accent recognition system's sensitivity in terms of how well it can distinguish between varieties which are much more similar. Past research ([1] [11]) has demonstrated Y-ACCDIST's ability to distinguish between the varieties in the *Accent and Identity on the Scottish/English Border* (AISEB) corpus [12].

AISEB was collected for sociolinguistic purposes and contains speakers from Berwick-upon-Tweed, Carlisle, Eyemouth and Gretna. A subset of AISEB was used to test Y-ACCDIST, where the recorded reading passage was taken from 30 speakers from each of the four locations in the corpus. Testing all 120 speakers in a *leave-one-out* training and testing configuration, Y-ACCDIST was able to classify these speakers into the four accent groups with a rate of 86.7% correct.

In a similar way to [9], [13] compared Y-ACCDIST against three different GMM-based accent recognisers. However, these comparative experiments were run on the AISEB corpus of geographically-proximate accents, rather than the ABI corpus. As already discussed, [9] found the text-dependent ACCDIST-based systems to outperform the text-independent GMM-based systems. [13] found that this was also the case when testing these kinds of systems on the AISEB corpus, but the ACCDIST-based systems appeared to be much more robust to more similar accent varieties. The text-independent GMM-based systems seemed to suffer much more due to the more challenging distinctions to make between more similar varieties. On the four-way AISEB accent classification task the GMM-UBM system achieved 37.5% correct, which is well below the 86.7% the Y-ACCDIST system achieved.

While researching system performance on geographically-proximate accents does move towards more forensically relevant experiments, there are of course still a number of aspects that are forensically relevant, which remain uncovered. The experiments presented below address other areas which are of interest to the forensic application: spontaneous speech data and degraded data.

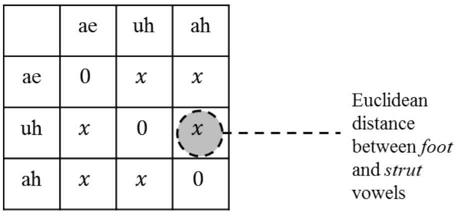
3. Experiments

First, this section gives a technical description of the inner workings of Y-ACCDIST. Following this, the experiments and results of testing Y-ACCDIST on conversational spontaneous speech data, and degraded data are given.

3.1. Y-ACCDIST Development Details

The following steps take place to train Y-ACCDIST and classify an unknown speech sample:

1. For each speaker in the training data, a speech sample and orthographic transcription are passed through a forced aligner. The forced aligner was built using the Hidden Markov Model Toolkit (HTK) [14] and a British English phoneset.
2. A midpoint MFCC (12 coefficients) was extracted for each phone in each speaker's sample.
3. An average MFCC is calculated for each phoneme in the phoneme inventory.
4. In a matrix (a Y-ACCDIST matrix), the Euclidean distance between every pair of phonemes is calculated using the average MFCC representations. One of these matrices is computed for each speaker. These distances are expected to capture accent-specific information. To illustrate, we can look at the *foot-strut split* in British English. A typical speaker of Northern English English will produce the vowels in *foot* and *strut* very similarly. A typical speaker of Southern English English, however, will produce the vowels of *foot* and *strut* differently. The Euclidean distance between these two vowels in a Northern speaker's matrix will therefore be smaller than the one calculated for a Southern speaker. An illustration of a matrix is given in the figure below:



	ae	uh	ah
ae	0	x	x
uh	x	0	x
ah	x	x	0

Euclidean distance between *foot* and *strut* vowels

Figure 1: Illustration of part of a Y-ACCDIST matrix.

For the experiments in this paper, all phonemes in the phoneset (vowels and consonants) were included in the Y-ACCDIST matrices

5. Using these Y-ACCDIST matrices to represent each of the training speakers, they are fed into a Support Vector Machine (SVM) with a linear kernel. For each accent group on rotation, the training speaker Y-ACCDIST matrices are plotted to form a *one-against-the-rest* in the SVM, which is effectively multi-dimensional space, forming an optimal 'hyperplane' each time. When classifying an unknown speaker's speech sample, it is first converted into a Y-ACCDIST matrix in the way described in steps 1-3. On each of the rotations for each accent, the unknown speaker's matrix is fed into the SVM. The accent group the unknown matrix forms the clearest margin with is the accent group the unknown speaker is assigned.

3.2. Spontaneous Speech

So far, Y-ACCDIST has largely been tested on tightly controlled experimental data. The speech data in past experiments on the AISEB corpus were done where each speaker was recorded reading the same passage. A change in dataset allows for more forensically relevant experiments to take place, where a large amount of spontaneous speech is produced by speakers of different accents. A description of the selected corpus is given below:

3.2.1. The Corpus

The data used for the experiments presented here were taken from the *Language Change in Northern English* corpus [15]. A subset of speakers' conversational speech recordings (with a sampling rate of 44.1kHz), along with their orthographic transcriptions were taken for these experiments. The speech recordings of 45 adult speakers (male and female) from Manchester, Newcastle and York (15 in each group), along with their orthographic transcriptions, were manually pre-processed. 10 minutes of net speech per speaker (and the accompanying orthographic transcriptions) were prepared for the experiments shown below.

3.2.2. Results

On the three-way accent classification task, distinguishing between speakers from Manchester, Newcastle and York, Y-ACCDIST achieved a recognition rate of 80.0% correct. Displayed below is the resulting confusion matrix for this task.

Table 1: Confusion matrix of results generated using spontaneous speech recordings.

Accent	Manc.	Newc.	York.
Manc.	**12**	0	3
Newc.	2	**12**	1
York.	1	2	**12**

3.3. Degraded Data

While testing Y-ACCDIST on spontaneous speech is more forensically relevant than previous experiments, the quality of the recordings is still unrealistic to the application. Forensic speech scientists mostly work with telephonic-quality or other degraded recordings. To begin to explore Y-ACCDIST's potential as an analytical tool on degraded data, the data used in the experiments above were degraded to a quality which resembles telephony. The recordings were downsampled to 8kHz, bandpass-filtered 250-3500Hz, and a-law compression was applied. The same experiments were re-run to achieve a recognition rate of 64.4% correct. The confusion matrix attached to this result is given in the table below:

Table 2: Confusion matrix of results generated using spontaneous speech recordings degraded down to telephonic quality.

Accent	Manc.	Newc.	York.
Manc.	**7**	4	4
Newc.	0	**13**	2
York.	2	4	**9**

Between the results generated from the good-quality data and those from degraded data, we can see an expected drop in performance. When we compare the two confusion matrices from each of the quality conditions, it is interesting to observe which particular accent groups appear to suffer more in the degraded condition. While the number of correctly classified Newcastle speakers seem to remain approximately the same under the degraded condition, the number of correctly classified Manchester and York speakers fall. This may be indicating that out of the three accent groups, Newcastle is the most distinct variety and so can withstand a more challenging data condition. Research on a larger data pool would be required to investigate this further.

3.4. Quantity of Data

One key criticism of the experiments presented above is the quantity of speech used to train and test the system. A 10-minute speech sample per speaker does not realistically align with what would normally be available to a forensic analyst. We can diminish the quantity of speech per speaker and run the system at different speech sample lengths. In increments of 30 seconds, the graph below demonstrates the effect of speech sample length on classification rate under the good-quality data condition.

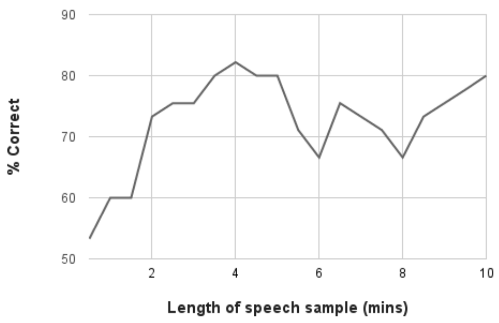


Figure 2: Classification rates with varying length of speech sample.

The graph shows a general improvement in performance between 30-second samples and 10-minute samples. However, it does not appear to be a stable increase in classification rate. While we reach a good recognition rate of 75.5% correct with 3 minutes of speech (a length much more reflective of what might be available to a forensic analyst), performance fluctuates beyond this. Due to the relatively small dataset, a larger dataset would be required to investigate this further. However, it might be that the change in segmental distribution that occurs in different lengths of speech samples has an unpredictable effect on performance.

4. Summary and Discussion

This paper has further entertained the idea of applying automatic accent recognition technology to forensic casework. The Y-ACCDIST system was applied to the *Language Change in Northern English* corpus to assess its performance on spontaneous speech data (compared to reading passage data which Y-ACCDIST has already been tested on). On a three-way accent classification task, 80.0% correct was achieved. Degrading this data down to a quality that resembled telephony generated a result of 64.4% correct.

These findings only just begin to explore a specific automatic accent recognition system's performance in a forensic context. Numerous avenues for further research exist. Below covers just three of these:

- As already stated above, much of the focus in forensic speech science is the task of speaker comparison. It would be interesting to see whether Y-ACCDIST could assist in some speaker comparison cases (i.e, can a Y-ACCDIST matrix represent a speaker's pronunciation pattern to specifically distinguish him or her from a wider speaker population?).
- Y-ACCDIST is highly reliant on segmental information to be able to work. In a given accent recognition task, certain segments are more telling than others (and this will vary depending on the particular accent varieties we are dealing with). It is therefore reasonable to assume that an unknown speech sample might need to contain a certain distribution of segments to obtain a reliable result. Another line of inquiry is to establish the segmental criteria an unknown speech sample needs to meet in order to be reliably processed and assessed by Y-ACCDIST.
- One important topic in forensic speech science is to do with the conclusion outputs of a system. In the experiments in this paper, only a closed-set classification task has been conducted. In reality, it might be more useful to determine the likelihood of a speaker belonging to a certain accent group with a more open-set approach. Integrating likelihood ratios is therefore considered an important development to the system.

Future research will target these directions.

5. References

[1] G. Brown, "Y-ACCDIST: An automatic accent recognition system for forensic applications," Master's thesis, University of York, UK, 2014.

[2] P. Foulkes and P. French, "Forensic speaker comparison: A linguistic-acoustic perspective," ser. The Oxford Handbook of Language and Law, P. Tiersma and L. Solan, Eds. Oxford University Press, 2012, pp. 557–572.

[3] P. Foulkes and K. Wilson, "Language analysis for the determination of origin," in *Proceedings of the 17th International Congress of Phonetic Sciences*, Hong Kong, 2011, pp. 692–694.

[4] M. Najafian, A. DeMarco, S. Cox, and M. Russell, "Unsupervised model selection for recognition of regional accented speech," in *Proceedings of Interspeech*, Singapore, 2014, pp. 2967–2971.

[5] M. Zissman, "Comparison of four approaches to automatic language identification of telephone speech," *IEEE Transactions on Speech and Audio Processing*, vol. 4, pp. 31–44, 1996.

[6] F. Biadsy, H. Soltau, L. Mangu, J. Navratil, and J. Hirschberg, "Discriminative phonotactics for dialect recognition using context-dependent phone classifiers," in *Proceedings of Odyssey: The Speaker and Language Recognition Workshop*, Brno, Czech Republic, 2010, pp. 263–270.

[7] T. Chen, C. Huang, E. Chang, and J. Wang, "Automatic accent identification using Gaussian Mixture Models," in *IEEE workshop on ASRU*, Italy, 2001.

[8] M. Huckvale, "ACCDIST: a metric for comparing speakers' accents," in *Proc. International Conference on Spoken Language Processing*, Jeju, Korea, 2004, pp. 29–32.

[9] A. Hanani, M. Russell, and M. Carey, "Human and computer recognition of regional accents and ethnic groups from British English speech," *Computer Speech and Language*, vol. 27, pp. 59–74, 2013.

[10] S. D'Arcy, M. Russell, S. Browning, and M. Tomlinson, "The Accents of the British Isles (ABI), corpus," in *Proceedings of Modelisations pour l'Identification des Langues*, Paris, France, 2004, pp. 115–119.

[11] G. Brown, "Automatic recognition of geographically-proximate accents using content-controlled and content-mismatched speech data," in *Proceedings of the 18th International Congress of Phonetic Sciences*, 2015.

[12] D. Watt, C. Llamas, and D. E. Johnson, "Sociolinguistic variation on the Scottish-English border," in *Sociolinguistics in Scotland*, R. Lawson, Ed. London: Palgrave Macmillan, 2014, pp. 79–102.

[13] G. Brown, "Automatic accent recognition systems and the effects of data on performance," in *Proceedings of Odyssey: The Speaker and Language Recognition Workshop*, Bilbau, Spain, 2016.

[14] 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*. Cambridge: Cambridge University Engineering Department, 2009.

[15] W. Haddican, P. Foulkes, V. Hughes, and H. Richards, "Interaction of social and linguistic constraints of two vowel changes in Northern England," *Language, Variation and Change*, vol. 25.

Holistic perception of voice quality matters more than L1 when judging speaker similarity in short stimuli

Eugenia San Segundo, Paul Foulkes, Vincent Hughes

Department of Language and Linguistic Science, University of York, UK

{eugenia.sansegundo|paul.foulkes|vincent.hughes}@york.ac.uk

Abstract

Spanish and English naïve listeners judged the similarity of 5 pairs of Spanish speaking identical twins. Listeners rated speaker similarity in a comparable way irrespective of their L1. This is of forensic relevance in non-native earwitness evidence, as it suggests that similar listening strategies operate (i.e. holistic approach to voice quality) when stimuli are short and no other segmental cues are available for the naïve listener – native or non-native – to judge speaker similarity.

Index Terms: voice quality, perception, twins, forensic phonetics, English, Spanish

1. Introduction

Voice quality (VQ) is defined as the quasi-permanent quality resulting from a combination of long-term laryngeal and supralaryngeal features, which typically makes a speaker's voice different from others [1]. The study of VQ has produced fruitful research in speech pathology and therapy [2], L2 phonology [3], and sociolinguistics, including studies exploring cross-dialectal patterns [4][5]. VQ serves as a social marker to indicate membership of a speech community [6], but it is also idiosyncratic. As such, it has received considerable attention in forensic phonetics, a discipline which applies phonetic knowledge to legal issues. Forensic Speaker Comparison (FSC) tasks, the most frequent in forensic casework [7], consist of the comparison of voice samples belonging to an offender and a suspect in order to assist courts in determining speaker identity.

The study of VQ can be approached from articulatory, acoustic or perceptual perspectives, including hybrid instrumental and perceptual assessment methods. In this investigation we focus on the auditory assessment of VQ, specifically as it is carried out by *naïve* listeners as opposed to *experts* (cf. technical speaker identification [8]). Our hypothesis is that under controlled conditions of speaker similarity (i.e. similar-sounding speakers sharing dialect, approximate age and mean F0), naïve listeners would rely on a holistic VQ perception in order to judge similarity between speakers. Native knowledge of the speaker language would be irrelevant when short stimuli belonging to different voices are very similar in segmental aspects. Only the combination of VQ characteristics (e.g. harsh voice, nasality or close jaw) would be available for the listener to judge speaker similarity. Under this holistic approach of VQ, both native and non-native listeners are expected to perform in a similar way.

To explain the factors which account for similarity ratings we nonetheless consider the perceptual evaluation carried out from a *featural* approach by a trained phonetician (expert listener). The holistic-featural dichotomy has traditionally accompanied the perceptual study of VQ and it continues to be an issue today. Previous studies [9][10] based on neuropsychological evidence suggest that the perception of VQ cannot be explained as the sum of separate features; instead, it involves a component of holistic, gestalt-like pattern processing. However, the different perceptual protocols (e.g. VPA; GRBAS; CAPE-V or SVEA; cf. summary in [11]) that are available for forensic phoneticians rely on the description of a voice in terms of a number of settings or perceptual dimensions: they are thus featural or componential analyses. How to handle the holistic-featural dichotomy is still a challenge, and more investigations are needed to explore in relation to how both perceptual approaches correlate or interact.

Laver's Vocal Profile Analysis protocol (VPA, [1]) is perhaps the most widely used analytic method whose components are referred to as 'settings', defined as long-term tendencies of the vocal apparatus to adopt a particular configuration [12][13]. Recent studies show growing interest in VQ – from an auditory perspective – for forensic purposes [14][15][16]; most using VPA or a simplified version of it. Despite the popularity of featural approaches, much remains to be explored as regards holistic judgments of voice quality made by lay listeners. This paper aims to fill this gap by looking at the role played by non-featural perception of VQ by naïve listeners, and to explore the language independence hypothesis of this holistic approach when judging speaker similarity. Some recent studies have explored lay perceptions of voice similarity, but without a focus on VQ. For instance, using Multidimensional Scaling (MDS) [17] proposes a method for assessing the degree of perceived similarity among a group of speakers for potential inclusion in voice parades. In [18] acoustic correlates are investigated for the perceptual dimensions obtained in the MDS analysis; and in [19] voice similarity judgements are found to depend on the accent background of the listener. Preliminary correlation results seem to show that different phonetic features contribute to the perceived similarity ratings for the two accents.

There are fewer studies focusing on the language dependency of VQ perceptual assessment. Most previous studies on native language effects in voice identification tend to suggest that native listeners have an advantage over non-natives [20][21]. Other investigations, however, fail to support this claim: [22] found that although identification improves the larger the phoneme repertoire in the voice sample, it is still possible to identify voices successfully when stimuli are random phonemes with no meaning and not belonging to any language. It can be then hypothesized that listeners pay attention to cues in a voice which do not require knowledge of the speaker's language, for instance suprasegmental aspects. Ho [23] found no native language effect when comparing British English and Chinese listeners in a speaker identification task where F0 was modified; listeners responded to the stimuli differently regardless of their L1, suggesting that F0 is a language-independent factor for voice identification.

In this study we focus on a different suprasegmental aspect (VQ), but the scope of the investigation differs from the above-mentioned studies in that we are not conducting identification tests or same-different tests. Instead, listeners are asked to rate speaker similarity, so that their ratings can be used as input to a MDS analysis in order to explore listeners' perceptual representations of very similar speakers. That is the reason why we selected a cohort of same-age, same-dialect speakers, with similar F0. Stimuli pairs belonging to monozygotic twins (MZ) were included as they represent extreme examples of voice similarity. Johnson & Azara [24] suggest that twins "serve as a unique control population for studies of the perception of personal identity". An important limitation of [24] is the heterogenous nature of the subjects (5 MZ twins and 1 dizygotic pair) with very different ages (20-67) and dialects. The first two dimensions of MDS solution in [24] correlated with age and dialect correspondingly. In our experiment a larger twin population is used, but most importantly age and dialect are controlled. Perceived speaker similarity is predicted to be explained solely by VQ characteristics, assessed holistically in a very similar way by native and non-native listeners.

2. Materials and method

2.1. Subjects

Five pairs of male MZ twins were selected from the corpus collected in [25]. All were native speakers of Standard Peninsular Spanish, and none reported any voice pathology. Three criteria were established in order to select only the most similar-sounding twin pairs from the corpus: (i) *similar age* (mean: 21, sd: 3.7); (ii) *similar mean F0* (mean: 113 Hz, sd: 13 Hz); and (iii) *similar Euclidean distance (ED)* between each speaker and his twin. EDs were based on the perceptual assessment of their VQ using a simplified version of the VPA scheme [26] by a trained phonetician (Author 1).

Table 1. *VPA speaker evaluation and Similarity Matching Coefficient (SMC) per twin pair. VT: vocal tract*

	VPA settings										
Speaker	**Labial**	**Mandibular**	**Lingual Tip**	**Lingual body**	**Velopharyngeal**	**Pharyngeal**	**Larynx Height**	**VT tension**	**Larynx tension**	**Phonation type**	**SMC**
AGF	0	0	0	0	0	0	2	1	1	1	
SGF	0	1	0	0	1	0	2	1	1	1	
Match	1	0	1	1	0	1	1	1	1	1	*0.8*
AMG	0	1	0	2	2	2	1	1	1	0	
EMG	0	1	0	2	0	2	1	1	0	0	
Match	1	1	1	1	0	1	1	1	0	1	*0.8*
ASM	1	0	0	0	1	0	2	0	1	1	
RSM	1	0	0	0	1	0	2	2	2	0	
Match	1	1	1	1	1	1	1	0	0	0	*0.7*
ARJ	2	2	1	1	0	1	2	2	1	0	
JRJ	1	2	0	0	0	0	2	2	2	0	
Match	0	1	0	0	1	0	1	1	0	1	*0.5*
DCT	0	2	0	2	0	2	0	2	1	1	
JCT	0	0	0	0	2	2	2	0	1	1	
Match	1	0	1	0	0	1	0	0	1	1	*0.5*

The simplified VPA consists in a reduced number of perceptual dimensions (10 settings) where each one is reduced to a continuum with three possibilities: neutral setting (labelled as 0) and two non-neutral possibilities (labelled as 1 and 2), typically going in opposite directions (e.g. labial setting: spread-neutral-round).

The EDs between twins were measured in Similarity Matching Coefficients (SMC), a typical distance measure for categorical data, where the number of matches for each variable is divided by the number of variables (Table 2). Mean SMC for all twins was 0.66, indicating that around 6 VQ settings were shared on average by the twin pairs.

2.2. Stimuli and listeners

2.2.1. *Stimuli*

Voice samples (~3 secs.) were extracted from semi-directed spontaneous conversations ([25]), held by the 10 twins individually with Author 1. The interlocutor is therefore controlled, resulting in the same type of speaking style in all conversations. All utterances were declarative sentences of different linguistic content (diverse neutral topics).

2.2.2. *Listeners*

Native Spanish speakers (N=20; age range 22-51, mean 33) and native English speakers with no knowledge of Spanish (N=20; age range 19-35, mean 25) took part in the perceptual experiment. They were recruited at universities in Spain and England, and none reported any hearing difficulty.

2.3. Design of perceptual test

A Multiple Forced Choice experiment was set up in Praat [26] with 90 different-speaker pairings, i.e. each speaker compared with everyone else. Stimuli were presented in random order and listeners had to indicate the degree of similarity of each stimuli pair on a scale 1 (very similar) to 5 (very different). Listeners were not told that the stimuli included twin pairs. The test was run on a PC with headphones in a silent room. A short pre-test with four voices (also twins but different ones) allowed familiarization with the test. Reaction times were measured from the end of the second stimulus. The test duration was approximately 15 minutes with a short break every 30 stimuli.

2.4. Analysis method

2.4.1. *Multidimensional Scaling*

The degree of perceived similarity was visualized using Multidimensional Scaling (MDS), a means of detecting meaningful underlying dimensions that explain observed similarities or dissimilarities (distances).

2.4.2. *Mixed-effects modelling*

Ordinal mixed effects modelling (MEM) was used to fit models to the similarity ratings using the *Ordinal* package in R [28]. The following fixed effects (predictors) were tested:

- Listener language – Spanish or English
- Similarity matching coefficient (SMC) – between the speakers in the target trial
- Reaction time
- Twins – whether speakers were twins or not

Random intercepts were fitted for listener and trial (i.e. target speaker comparison). The first model tested for the effect of language on the similarity ratings provided by listeners. A step-up approach was then adopted whereby predictors and interactions were added iteratively and models compared using ANOVAs. Model comparison was conducted in order to assess the best fit to the data.

3. Results

3.1.1. Multidimensional Scaling

MDS analyses were conducted using the similarity scores. The relative magnitude of the sorted Eigenvalues indicates that seven dimensions would be necessary to accurately reproduce between-speaker distances in the perceptual space for both English and Spanish listeners (stress: 0.03 for Spanish listeners; 0.07 for English listeners). However, MDS results are typically visualized using only the first 2 or 3 dimensions. Figures 1-2 show MDS plots for Spanish and English listeners respectively using 2 dimensions (stress: 0.8). Each point represents the location of a speaker in the listeners' perceptual space. 3D models showed an important drop in stress (0.4). Table 2 shows the normalized intra-pair Euclidean distances taking into account the seven dimensions in which listeners' ratings seem to be based.

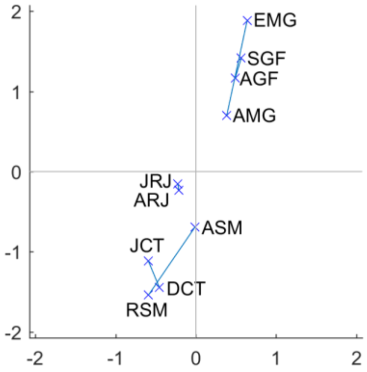


Figure 1: *MDS 2D plot (Spanish listeners)*

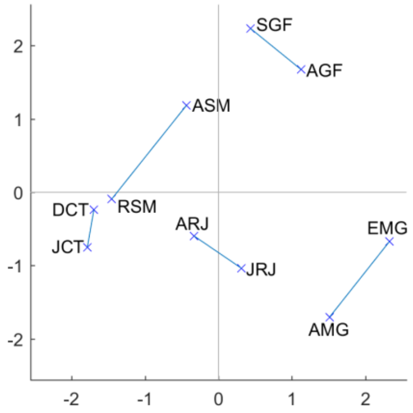


Figure 2: *MDS 2D plot (English listeners)*

Table 2: *Normalized intra-pair Euclidean distances based on seven perceptual dimensions*

speakers →	AGF	DCT	ARJ	ASM	AMG
listeners ↓	SGF	JCT	JRJ	RSM	EMG
Spanish	0.341	0.343	0.345	0.369	0.607
English	0.264	0.219	0.349	0.435	0.445

3.1.2. Mixed-effects modelling

The best model fitted to the data based on model comparison incorporated all fixed effects and interactions between fixed effects. Significant interactions were found between language and both reaction time and whether the target trial contained a twin pair or not. For English listeners, similarity ratings were not affected by reaction time. That is, listeners were no more likely to judge speaker pairs as being very similar or dissimilar as a function of reaction time. However, Spanish listeners were more likely to give a rating of 5 ('very dissimilar') if reaction time was short, and a rating of 1 ('very similar') if reaction time was longer. This is of special interest if we consider that average reaction times were very similar for Spanish (mean: 0.82 secs; SD: 0.14) and English listeners (mean: 0.84; SD: 0.18).

A number of language independent effects were also found. Across all listeners, twin pairs were rated as being more similar to each other (ratings closer to 1) than non-twin pairs. Twin pairs with low SMC values (i.e. those who are objectively more similar to each other) were also rated as being more similar than twin pairs with high SMC values. For non-twin pairs, listeners did not rate speaker pairs with higher or lower SMC values as being more or less similar. Finally, for twin pairs all listeners were more likely to respond with 1 (i.e. very similar) if reaction time was short. Conversely, for non-twin pairs, all listeners are more likely to respond with 5 if they respond quickly. If they did not respond quickly, reaction time was no predictor of similarity rating.

4. Discussion

MDS analyses show that the optimal configuration to visualize speaker distances would require a 7-dimensional space (lowest possible positive stress value). 2D plots are therefore poorer representation of the data, reflected in a high stress value. This confirms what is well known for VQ: its high multidimensionality. A thorough understanding of perceptual judgements by naïve listeners require other types of analyses, and that is why MEM was conducted. Even though we cannot explain listener decisions with only two dimensions, we still find similar trends in both listener groups, like extreme closeness of speakers DCT and RSM. When we look at their featural VQ analysis (Table 1), the only setting that they share relates to vocal tract tension, possibly evidencing the higher salience of this setting.

Since seven dimensions seem to best represent listeners' perceptual space, we ordered twin pairs from most to least similar and the same ranking appears in both listener groups: AGF-SGF and DCT-JCT being the closest twin pair with slight differences in their normalized intra-pair Euclidean distances; ASM-RSM and AMG-EMG consistently appearing as the least similar pairs for both listener groups. Looking in detail at Table 1, we find that settings shared by AGF-SGF and DCT-JCT relate to the larynx (laryngeal tension and phonation types). Supralaryngeal matches (e.g. labial and lingual tip) are due to shared neutral settings, which should probably not weigh in the same way as matches due to deviations from neutrality in future SMC calculations. In contrast, matches for ASM-RSM and AMG-EMG (discarding matches based on shared neutral settings), are only supralaryngeal matches. These seem not to be so salient for naïve listeners from a holistic perspective, as these twin pairs are consistently far apart in the listeners' perceptual space. This finding seems to point to the same cue prominence by all naïve listeners, i.e. regardless of language familiarity or understanding of the linguistic content.

Table of Contents for this manuscript

Equivalent reaction times suggest similar listening strategies ('gut' impressions; holistic VQ perception). However, qualitative feedback from participants suggest that other cues, mostly rhythmic aspects (e.g. speaking rate) may have contributed to perceived similarity as well. These deserve future investigations, as they are also suprasegmental features, apparently also salient even in short stimuli and possibly independent of the listener L1.

Mixed effects modelling revealed a number of effects involving language, although no clear indication of different listening strategies across groups. Significant effects involved reaction time, indicating that, for certain target pairs, similarity ratings are different for English and Spanish listeners depending on how long it took to make the decision. Notably, statistical modelling did suggest a number of language independent factors. Most notably, twin pairs were rated as being more similar to each other than non-twin pairs irrespective of listener language.

5. Conclusions

It is well known that multiple factors affect unfamiliar naïve recognition, from individual listener ability to the distinctiveness of the speaker's voice; the contribution of the latter not being fully understood. In order to explore in which VQ aspects speaker distinctiveness may lie, we have designed a perceptual test where Spanish and English listeners had to rate speaker similarity in pairwise comparisons. Results have shown that when other linguistic cues are suppressed –because of short stimuli– native and non-native listeners rate speaker similarity in a very similar way. Using short speech samples makes it difficult for listeners to base their similarity judgments in other aspects which are not VQ. Although these results should not be extrapolated to earwitness evidence with different characteristics and the native advantage may still hold true in situations where listeners are exposed to longer speech samples, this investigation has aimed to explore the role of VQ holistic perception, which seems to be the resource available for lay listeners to judge speaker similarity at least in a homogenous population of same-accent, same-age, similar-sounding speakers. Future investigations will look further at interrelationships between naïve holistic VQ and the featural decomposition of VQ by expert listeners, as the latter reveals speaker similarities for specific settings that do not appear to be salient in the holistic perception of VQ or at least do not have a strong weight in the similarity ratings made by naïve listeners.

6. Acknowledgements

This research was funded via the UK AHRC grant *Voice and Identity – Source, Filter, Biometric* (AH/M003396/1). We thank Olaf Köster and Jose Antonio Mompean for their helpful insights as well as Juana Gil for her assistance in the recruitment of Spanish listeners.

7. References

[1] Laver, J. The Phonetic Description of Voice Quality, Cambridge University Press, 1980.

[2] Webb, A. L., Carding, P. N., Deary, I. J., Mackenzie , K., Steen, N., Wilson, J. A. "The reliability of three perceptual evaluation scales for dysphonia", European Archives of Otorhinolaryngology, 261:429–434, 2004.

[3] Esling, J. H. & Wong, R.F. "Voice quality settings and the teaching of pronunciation", TESOL Quarterly 17:89–95, 1983.

[4] Stuart-Smith, J. "Glasgow: accent and voice quality", in P. Foulkes & G. Docherty [Eds], Urban Voices, 203-222, Arnold, 1999.

[5] Esling, J. H. Voice quality in Edinburgh: a sociolinguistic and phonetic study. PhD dissertation, University of Edinburgh, 1978.

[6] Laver, J. Principles of Phonetics, CUP, 1994.

[7] Foulkes, P., & French, P. "Forensic speaker comparison: a linguistic-acoustic perspective", in L. Solan & P. Tiersma [Eds], The Oxford Handbook of Language and Law, 418-421, Oxford University Press, 2012.

[8] Rose, P. "Forensic speaker identification", CRC Press, 2003.

[9] Kreiman, J. & Gerrat, B. "Comparing two methods for reducing variability in voice quality measurements", Journal of Speech, Language and Hearing Research, 54:803-812, 2011.

[10] Kreiman, J. & Sidtis, D. Foundations of Voice Studies, Wiley-Blackwell, 2011.

[11] Gil Fernández, J. & San Segundo, E. "La cualidad de voz en fonética judicial", in E. Garayzábal, M. Jiménez & M. Reigosa [Eds], Lingüística forense. La lingüística en el ámbito legal y policial, 154-199, Euphonía Ediciones, 2013.

[12] Honikman, B. "Articulatory settings", in D. Abercrombie, D.B. Fry, P.A.D. MacCarthy, N.C. Scott, & J.L.M. Trim [Eds], In Honour of Daniel Jones, 73–84, Longman, 1964.

[13] Beck, J. "Perceptual analysis of voice quality: the place of Vocal Profile Analysis", in W.J. Hardcastle & J. Mackenzie Beck [Eds], A Figure of Speech: a Festschrift for John Laver, 285–322, Laurence Erlbaum Associates, 2005.

[14] González-Rodríguez, J., Gil, J., Pérez, R., & Franco-Pedroso, J. "What are we missing with i-vectors? A perceptual analysis of i-vector based falsely accepted trials", Proc. Odyssey 2014, 33-40, 2014.

[15] French, P., Foulkes, P., Harrison, P., Hughes, V., San Segundo, E. & Stevens, L. "The vocal tract as a biometric: output measures, interrelationships, and efficacy", Proc. 18th ICPhS, Glasgow, 2015.

[16] San Segundo, E., Hughes, V., French, P., Foulkes, P. & Harrison, P. "Developing the vocal profile analysis scheme for forensic voice comparison", BAAP Colloquium, Lancaster, 2016.

[17] McDougall, K. "Assessing perceived voice similarity using multidimensional scaling for the construction of voice parades", Int. Journal of Speech Language and the Law, 20:163-172, 2013.

[18] Nolan, F., McDougall, K. & Hudson, T. "Some acoustic correlates of perceived (dis)similarity between same-accent voices", Proc. 17th ICPhS, Hong Kong, 2011.

[19] McDougall, K., Hudson, T. & Atkinson, N. "Perceived voice similarity in Standard Southern British English and York English" UKLVC Conference, York, 2015.

[20] Perrachione, T.K., Pierrehumbert, J.B. & Wong, P.C.M. "Differential neural contributions to native- and foreign-language talker identification", Journal of Experimental Psychology: Human Perception and Performance, 35:1950-1960, 2009.

[21] Köster, O. & Schiller, N. O. "Different influences of the native language of a listener on speaker recognition", Forensic Linguistics, 4:8-28, 1997.

[22] Bricker, P. D., & Pruzansky, S. "Effects of stimulus content and duration on talker identification", Journal of the Acoustical Society of America, 40:1441-1449, 1966.

[23] Ho, C.-T. Is pitch a language-independent factor in forensic speaker identification?, MA diss., University of York, 2007.

[24] Johnson, K. & Azara, M. "The perception of personal identity in speech: evidence from the perception of twins' speech" Unpublished manuscript, 2000.

[25] San Segundo, E. Forensic speaker comparison of Spanish twins and non-twin siblings, PhD dissertation, Menéndez Pelayo International University & CSIC, 2014.

[26] San Segundo, E. & Mompean, J. "Voice quality similarity based on a simplified version of the Vocal Profile Analysis: a preliminary approach with Spanish speakers including identical twin pairs", Sociolinguistics Symposium 21, Murcia, 2016.

[27] Boersma, P., & Weenink, D. Praat: doing phonetics by computer [Computer software] (Version 5.3.79), 2012.

[28] Christensen, R.H.B., "Ordinal", R Package [Computer software] (v.3.3.0), 015.

Lexical manipulation as a discovery tool for psycholinguistic research

Laurence Bruggeman[1,2] *and Anne Cutler*[2]

[1]Department of Linguistics, Macquarie University, Australia
[2]The MARCS Institute, Western Sydney University, Australia

laurence.bruggeman@mq.edu.au; a.cutler@westernsydney.edu.au

Abstract

Consultation of machine-readable dictionaries has advanced understanding of language processing; but these resources also allow examination of processing consequences if the lexicon changes. To recognise speech, listeners must rapidly evaluate spoken input as matching or mismatching candidate words. Listeners use any speech cues that help this process, whereby identical cues across languages may be used in one language but not in another. Suprasegmental stress cues, for example, are similar in Dutch and English, but used only in Dutch. This asymmetry has been explained as due to vowel reduction in English; lexical manipulation here tests this proposal and suggests a refinement.

Index Terms: lexicon,stress,word recognition, English, Dutch

1. Introduction

Electronic versions of complete dictionaries have now been available to language researchers for more than three decades. Their availability revolutionised both the design of automatic speech recognition systems and the modelling of language processing by human speakers and listeners. For example, an entire class of early spoken-word recognition models [1,2] that was based on strictly sequential recognition of words in their order of arrival was rendered untenable by the discovery that most English words cannot be uniquely recognised until at or after their final sounds [3]; only the automatically searchable dictionary resources made the latter discovery possible.

In the intervening decades, psycholinguists have made good use of such dictionary tools, which have become available for a steadily increasing number of languages. The structure of vocabularies is well understood and differences between vocabularies have been easy to measure, enabling their processing implications to be derived and tested. Thus the makeup of language phoneme repertoires determines average word length in dictionaries of that language [4] and predicts that this will carry through to everyday speech experience, as is confirmed in standard spoken samples [5].

However, electronic dictionaries offer researchers the opportunity of going beyond the usual activities involved in dictionary use – consulting re a single item, searching for the full extent of classes of items, analysing overall structure. Electronic dictionaries can also be deliberately altered.

It may seem as if altering a dictionary would be a pointless exercise; it can, after all, have no effect on how users of a language actually deploy their vocabulary resources. But lexical manipulation in fact provides a tractable and effective way of testing psycholinguistic explanations or predictions in which vocabulary structure plays an operational role. Section 2 describes two case studies of research in which the lexical manipulation method was used to illuminate differing questions, and in Section 3 we apply the technique to a new case involving a cross-language processing asymmetry.

2. Lexical Manipulation

2.1. Phoneme substitution

Lexical manipulation can shed light on effects of phonemic confusion, in particular distinctions that prove intractable for second-language listeners (such as the English contrast in, e.g., *write*, *light* for many listeners with an Asian native language, or in *cattle*, *kettle* for listeners whose first language is Dutch or German). These two phoneme confusions were at issue in [6], in which the English lexicon was examined from the point of view (or hearing) of second-language users for whom these distinctions were not perceptible in word-initial or word-medial position.

The manipulation in [6] thus involved treating pairs like *write*/*light* or *cattle*/*kettle* as homophones, and replacing [r] by [l] and [æ] by [ɛ], consistent with independent evidence of the direction of these mergers [7,8]. The results revealed the greatest effect of phonemic confusion to be located not in whole-word substitutions (despite their salience to language users) but in increase in the two major measures of inter-word competition: embedding (spuriously, *egg* in *agriculture*, *let* in *reticent*) and overlap (*lemon* and *remedy* or *matter* and *metal* are heard as starting with the same syllable). Moreover, the effect of a consonantal confusion was significantly greater than that of a vowel confusion, and confusion direction was asymmetric: listeners with [r]/[l] difficulty would actually have less trouble collapsing English [r] and [l] to [r], rather than to [l]! These findings suggest lines of speech perception research and also possible avenues for listening training.

2.2. Phonological rule substitution

The lexical manipulation technique has also been deployed to elucidate different patterns of embedded-word location across languages [9]. These analyses (without lexical manipulation) showed that the predominance of word-initial over word-final location of embedded words (such as *cat* or *log* in *catalogue*), long known for English, was replicated even more strongly in the related languages Dutch and German, but did not appear in Japanese. Japanese has neither suffixes nor stress, while all the other languages have both. The French vocabulary, which has one of those two features (suffixes) but not the other (stress), fell in between the Japanese and English values, thus suggesting that both factors played a role. This was tested in [10] by a radical form of lexical manipulation, which has been termed “lexical miscegenation”, since it effectively creates a hybrid of separate varieties. The French lexicon was augmented with pronounced schwa in all possible legal positions (giving *fille* two syllables and *petite* three). Its embedding pattern then came to closely resemble that of English and other stress languages. The contribution of stress to embedding patterns is thus based on schwa distributions.

3. Explaining asymmetry in speech cue use

3.1. Speech cue use

It is a remarkable fact that listeners do not always make optimal use of speech cues in recognizing spoken utterances. A striking example in phoneme recognition concerns the use of transitional cues to identify fricative sounds, which only happens when fricatives must be distinguished from highly similar alternatives [11,12]. If a sound has no close competitor sounds, available information in the signal may prove redundant and thus be ignored. The explanation of this pattern thus invokes language-specific phoneme repertoires.

At the word recognition level, similar asymmetries are found. The process of word recognition involves sorting out the intended words in a heard utterance from among the alternative candidate words that are embedded (i.e., fully supported) or overlapping (i.e., partially supported) in the speech signal. If there is one thing that is well known about spoken-word recognition, it is that listeners do not wait around to hear utterances or words or even syllables in full before attempting to recognize them; they constantly consider alternative interpretations and weigh the continually incoming evidence in the signal in terms of whether it offers support for, or counts against, the current options. Spoken *Speech Science and Technology* may activate the embedded words *pea/bee*, *peach/beach, each*, *sigh*, and *knowledge*, the cross-word embeddings *sand* or *antic*, and many temporarily supported candidates such as *speed*, *sign*, *text*, or *echo*. Such words are effortlessly discarded by listeners, though traces of their temporary presence can be discerned in psycholinguistic experiments. For instance, in the cross-modal priming task, where listeners make yes-no lexical decisions about visually presented words while hearing speech, words are recognised more slowly when they partially match the auditory input than when the auditory input is totally unrelated (e.g., responses to visual *feel* are slower after spoken *feed* than after spoken *name*; [13]). This response inhibition represents the temporary availability but later rejection of an alternative interpretation (*feel*) of the speech input. Because the word has been rejected it is momentarily less available to the lexical decision process.

With this same task, cross-language differences have appeared in the use of stress cues in spoken-word recognition. In English, for instance, there are minimal pairs of words that differ only in stress, such as *INsight*/*inCITE* (henceforth, upper case denotes a primary-stressed syllable). Dutch also has such minimal pairs, e.g., *VOORnaam* 'first name' vs. *voorNAAM* 'respectable'. English and Dutch are closely related languages, both having free lexical stress realised under highly similar rules [14]. Yet in cross-modal priming, if the mismatch involves only stress, English listeners do not show the inhibition that indicates a word candidate's rejection. Thus the spoken initial part of *admiRAtion*, *admi-*, does not inhibit responses to visual *admiral* (initial stress: *ADmiral*; [15]). In Dutch, in contrast, such mismatch indeed leads to inhibition – e.g., responses to visually presented *dominee* 'pastor' (pronounced *DOminee*) are slowed after *domi-* from *domiNANT* 'dominant'[16]. It seems that in these two very similar languages, in which stress is realised with virtually identical suprasegmental cues, Dutch listeners use the cues but English listeners don't (Fig. 1). The English result is only one of many demonstrations (in multiple tasks: recognition in noise [17]; acceptability judgment [18]; phoneme detection [19]; goodness rating [20]) of English listeners ignoring suprasegmental cues to word identity.

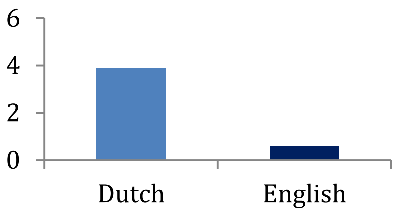


Figure 1 (data from [16] and [15]): *Dutch and English* cross-*modal priming: Inhibition due to mismatching stress (RT given mismatching prime minus RT given control prime, as % of control condition RT). The effect is highly significant in Dutch, insignificant in English.*

3.2. Lexical asymmetry as explanation

Lexical statistics suggest why English listeners should behave in this way. An estimate of the strength of competition from co-activated candidate words in speech recognition can be derived by tallying word embedding and overlap in the vocabulary, and this can be done taking only segmental structure into account, or stress pattern as well. In *enterprise*, for instance, *enter* and *prize* are embedded, and *settee* has *set* and *tea*. But if we require primary stress location to match also, *ENterprise* contains only *enter*, and *setTEE* contains only *tea* (neither *set* nor *prize* have primary stress there). Dutch examples are *OUderdom* 'old age' (*ouder* 'parent' and *dom* 'stupid', or only *ouder*) and *karPET* 'rug' (*kar* 'cart' and *pet* 'cap', or just *pet*). Dutch embedding competition reduces by more than 50% if stress match is indeed required in this computation [21]. There is a smaller (but also significant) reduction in English. But if the numbers are weighted by carrier word frequency to estimate actual competition in natural speech, a Dutch segments-only count gives 1.52 competitors per word of speech on average, while a segments-plus-stress count reduces this to .74. The equivalent numbers for English are .94 and .59 [21]; see Fig. 2. This is a quantum improvement for Dutch (from more than one to under one) but not for English (under one to a bit further under one).

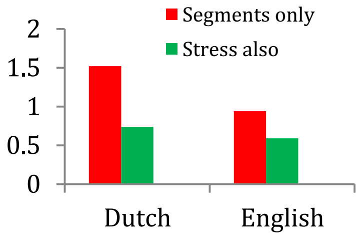


Figure 2 (data from [21]): *Dutch and English: Estimated number of competing words per word of speech, given match in segments only, or in segments and stress.*

One possible explanation lies in the fact that English syllables *without* primary stress often have a particular vowel, namely schwa. English listeners need attend only to segments in computing mismatch to competitors, as suprasegmental cues do not reduce competition sufficiently to make recourse to suprasegmental processing mechanisms worth the effort.

3.3. Dutch and English statistics

We test the hypothesis that schwa distribution might be the cause of this asymmetry (as it was the cause of the embedding asymmetry examined in [10]), by analysing the CELEX database [22]. This contains both English (British version) and Dutch lexicons. These closely related languages share many phonological attributes (e.g., both allow complex onsets and codas and have short vowels as well as long vowels and diphthongs). The phoneme inventory of the two languages differs somewhat; English has more consonants than Dutch and also more vowels, but each is a member of the class of languages with a rather high number of both phoneme types.

CELEX's lexicon of Dutch, as it happens, is much larger than its English lexicon. The Dutch lexicon contains more than 124,000 lemmas, and frequency statistics based on a 42 million word corpus; the English lexicon contains 52,000 lemmas, and frequencies from a 17.9 million word corpus. These differences – in the size of the frequency count corpora especially – largely arise from differences in the original sources from which the database was compiled. However, the differences in lexicon size also in part reflect cross-language morphological differences. Because of the size differences, absolute totals are obviously not directly comparable. Therefore our calculations are based on proportions.

We first assessed whether Dutch and English in fact differ in the overall frequency of schwa. Both languages have stress, both have vowel reduction, and both have extensive affixation with affixes (both suffixes, especially inflectional suffixes, and prefixes) typically realised by weak syllables containg schwa. Here the morphological differences between the two languages would in fact tend to bias the frequency of schwa towards more in Dutch, since Dutch word formation both uses prefixing more extensively than English (e.g., *gevoel* 'feeling', *geval* 'case', *geluk* 'luck' and many more, with the syllable *ge-* containing schwa) and includes a larger range of verbal inflexions including both prefixes and suffixes (thus where English *walk* adds the forms *walks*, *walking* and *walked*, Dutch *lopen* 'walk' adds *loopt*, *loop*, *liep, liepen*, *gelopen, lopend, geloop*). We thus initially computed the proportion of full vowels versus schwa in different word positions in each lexicon. For this count we included lemmas (base word entries) only, since including all word forms, especially in the case of verbs, leads to a greater size asymmetry between the lexicons and, for the reasons listed above, hugely increases the number of syllables with schwa, particularly in final position. Using lemmas still leaves morphologically caused asymmetries, since in fact all verb infinitives of Dutch (e.g., *lopen*) are marked as such by a final *–en* (with schwa). Similarly, a very large number of Dutch nouns derived from verbs begin with the prefix *ge-* which is the past participle inflection (e.g., *gebouw* 'building').

Indeed, even this lemma-based count revealed that Dutch final syllables are somewhat more likely to contain schwa than English final syllables, and while the proportions for initial syllables were fairly close for the two languages, there was a slight predominance of the schwa count in Dutch there too. Thus neither in initial nor in final syllables did we find support for greater frequency of schwa in English.

Medial syllables, however, patterned differently. Here the languages tend towards opposite patterns, as Figure 3 shows. In all word lengths with medial unstressed syllables (3 syllables and above), Dutch has more full vowels in such syllables (and English has more schwa).

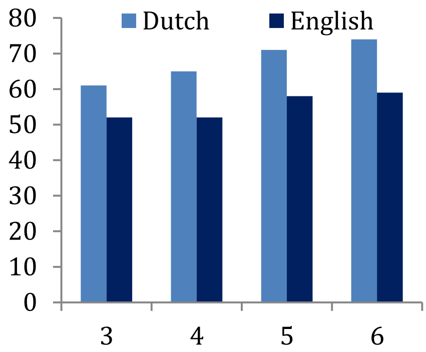


Figure 3: *Dutch and English: Percent word-medial syllables containing a full vowel, for words of 3 to 6 syllables. The equivalent figure for word-medial syllables containing schwa is the reverse of this (English always the higher value).*

We next assessed whether Dutch and English differ in the relative amount of competition offered by the makeup of their respective vocabularies. We tallied for each phoneme, the number of words beginning with that phoneme; for each syllable, the number of words beginning with that syllable; and for each bisyllabic string, the number of words beginning with that bisyllable. Given the size asymmetry between the two CELEX lexicons, it was necessary to adjust these results for the relative number of words involved. For the phoneme measure and the first-syllable measure we found no significnt cross-vocabulary difference. Thus there is (as expected) no tendency for either language to have, in principle, more inter-word similarity and hence a greater amount of competition. The two-syllable measure revealed that English in fact had more such overlap in its vocabulary. This was indubitably due to the prevalence of schwa; consider that English *coral*, *correlate*, *corridor*, *coroner*, *corrugated* and *coryphee*, despite being all spelled differently, begin in each case with the same CVCV sequence, the constant second V being schwa.

3.4. Stress cue substitution

The calculations so far do not suggest that Dutch listeners use suprasegmental cues to recognise words because their lexicon confronts them with more onset-overlapping competitor pairs; on the contrary, English presents the more difficult listening task in this respect. This renders the next step in this project even more interesting. We now undertake some lexical manipulation in order to shed further light on this issue; the manipulation involves moving one of these two lexicons in the direction of the other (a form of miscegenation, as in [10]).

Here we can either choose to make Dutch more English-like (having a greater proportion of schwa, especially in medial position), or English more Dutch-like (having a greater proportion of full vowels in medial position) – or indeed both, though such duplication would of course deliver no additional information. The replacement of full vowels by schwa, a many-to-one mapping, is the simpler calculation. The replacement of schwa by full vowels, in contrast, is a one-to-many mapping, so inherently more difficult; furthermore, it raises difficult issues of phonological principle in the choice of mapping (e.g., should the schwa in *coral, correlate*, *corridor*, etc. be replaced by the same full vowel, or by different ones?). For these reasons we chose the simpler option and undertook an alteration of the CELEX Dutch lexicon.

Table of Contents for this manuscript

To obtain an estimate of embedding frequency, we altered the proportions of schwa/full vowel in polysyllabic Dutch words to bring the schwa distribution in this lexicon in line with that in the English one, and computed embedding statistics analogous to those reported by [21] and summarised in Figure 2. Compared with the unmodified Dutch lexicon, embedding in this altered lexicon was reduced by almost 10%. To further obtain a measure of overlap competition, we selected the two-syllable overlap set computed for Dutch and described above (section 3.3), and compared it across the original unmodified lexicon and the altered lexicon. Again, this led to a reduction of competition in the latter, albeit this time only by 3% if stress location was ignored and by 2.5% if competitors had to match in the location of primary stress.

These two consistent outcomes indicate that the Dutch lexicon would contain fewer competitors if more use was made of the optional reduction to schwa provided by the phonological rules of the language; by implication there is thus support for the use of suprasegmental stress cues where they can assist in narrowing lexical competitor sets. This is in line with one part of the proposed explanation of why Dutch listeners use such stress cues in spoken-word recognition while English listeners ignore them, though it does not directly address English listeners' choices. The English lexicon, in part because of the predominance of schwa in medial syllables, confronts listeners with more competition. Schwa in English or Dutch cannot bear stress, and as a result of this, English listeners actually have very few opportunities to profit from suprasegmental stress cues to resolve competition.

4. Conclusions

The structure of vocabularies determines both what listeners have to recognise to understand spoken utterances (stand-alone words, agglutinative particles, polysynthetic sequences?) and also what speech cues are necessary and worthwhile to attend to in the task. In the case of suprasegmental cues to the distinction between primary and secondary stress, abundant empirical evidence shows that Dutch listeners make good use of the information they provide, but English listeners do not. The cues are acoustically similar in both languages, and Dutch listeners are capable of using the English cues even to the point of outdoing native listeners in an identification task [15], but English listeners fail to attend to them [17-20].

Lexical manipulation has supported the suggestion that English vowel reduction plays a role in this pattern. However, our results overall counter-indicate a larger claim that vowel reduction alone is responsible, because overall, and in initial or final syllables, Dutch has a greater proportion of syllables with schwa than English. More can be discovered here, and in particular we note again the higher proportion of morphological affixes in the Dutch lexicon. CELEX does not provide sufficient information for us to calculate exactly the relative proportion of affixes with schwa, but it does indicate an upper and a lower bound: English has a lower proportion of affixes with schwa in the lexicon than Dutch does (75% of the Dutch level at most, but at least 23.5%). Future work could address the processing implications of this asymmetry and its potential role in influencing word recognition processes.

We strongly encourage other researchers to join us in this view of dictionary resources: dictionaries are not holy untouchables, they are alterable objects that can serve as tools to discover novel potential patterns and hence to prompt new and innovative empirical studies of language processing.

5. Acknowledgements

Support is acknowledged from the ARC Centres of Excellence in Cognition and its Disorders (LB) and for the Dynamics of Language (AC) as well as from the Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands (AC, LB).

6. References

[1] Cole, R. and Jakimik, J., "Understanding speech: How words are heard", in G. Underwood [Ed.], Strategies of information processing, 67-116, Academic Press, 1978.

[2] Marslen-Wilson, W. and Welsh, A., "Processing interactions and lexical access during word recognition in continuous speech", Cog. Psychol., 10: 29-63, 1978.

[3] Luce, P., "A computational analysis of uniqueness points in auditory word recognition", Perc. Psychophys., 39: 155-158, 1986.

[4] Cutler, A., Norris, D. and Sebastian-Galles, N., "Phonemic repertoire and similarity within the vocabulary. Proc. INTERSPEECH 2004, Jeju, Korea; vol. 1, 65-68, 2004.

[5] Maddieson, I., "Word length is (in part) predicted by phoneme inventory size and syllable structure", J. Acoust. Soc. Am., 132: 2218, 2016.

[6] Cutler, A., "The lexical statistics of word recognition problems caused by L2 phonetic confusion", Proc. INTERSPEECH 2005, Lisbon, 413-416, 2005.

[7] Iverson, P. et al., "A perceptual interference account of acquisition difficulties for non-native phonemes", Cogn., 87, B47–B57, 2003.

[8] Broersma, M., *"Kettle* hinders *cat*, *shadow* does not hinder *shed*: activation of 'almost embedded' words in nonnative listening", Proc. INTERSPEECH 2007, 1893-1896, 2007.

[9] Cutler, A., Otake, T. and Bruggeman, L., "Phonologically determined asymmetries in vocabulary structure across languages", J. Acoust. Soc. Am., 132: EL155-EL160, 2012.

[10] Cutler, A. and Bruggeman, L., "Vocabulary structure and spoken-word recognition: Evidence from French reveals the source of embedding asymmetry", Proc. INTERSPEECH 2013, Lyon, 2812-2816, 2013.

[11] 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.

[12] Wagner, A., "Cross-language similarities and differences in the uptake of place information", J. Acoust. Soc. Am., 133: 4256-4267, 2013.

[13] Marslen-Wilson, W., "Activation, competition, and frequency in lexical access", in G.T.M. Altmann [Ed.], Cognitive Models of Speech Processing, 148–172, MIT Press, 1990.

[14] Trommelen M. and Zonneveld, W., "Word stress in the West Germanic languages", in H. van der Hulst [Ed.], Word Prosodic Systems in the Languages of Europe, 478-515, Mouton, 1999.

[15] Cooper, N., Cutler, A. and Wales R., "Constraints of lexical stress on lexical access in English: Evidence from native and non-native listeners", Lang. Speech, 45: 207–228, 2002.

[16] Van Donselaar W., Koster, M. and Cutler, A., "Exploring the role of lexical stress in lexical recognition", Quart. J. Exp. Psy., 58A: 251–273, 2005.

[17] Slowiaczek, L., "Effects of lexical stress in auditory word recognition", Lang. Speech, 33: 47-68, 1990.

[18] Slowiaczek, L., "Stress and context in auditory word recognition", J. Psycholing. Res., 20:465-481, 1991.

[19] Small, L., Simon, S. and Goldberg, J.S., "Lexical stress and lexical access: Homographs versus nonhomographs", Perc. Psychophys., 44:272-280, 1988.

[20] Fear, B., Cutler, A. and Butterfield, S., "The strong/weak syllable distinction in English", J. Acoust. Soc. Am., 97:1893-1904, 1995.

[21] Cutler A. and Pasveer D., "Cross-linguistic differences in effects of lexical stress on spoken-word recognition", Proc. 3rd Int. Conf. Speech Prosody, Dresden, 237–240, 2006.

[22] Baayen, H., Piepenbrock, R. and van Rijn, H., The CELEX lexical database. Philadelphia: Linguistic Data Consortium, University of Pennsylvania (CD-ROM), 1993.

Sub-band cepstral variability within and between speakers under microphone and mobile conditions: A preliminary investigation

Frantz Clermont[1,2], *Yuko Kinoshita*[3], *Takashi Osanai*[4]

[1]Prince Sultan University, Saudi Arabia
[2]J.P. French Associates Forensic Lab., UK
[3]College of Arts and Social Science, The Australian National University, Australia
[4]National Research Institute of Police Science, Japan

dr.fclermont@gmail.com, yuko.kinoshita@anu.edu.au, osanai@nrips.go.jp

Abstract

We describe preliminary results of a sub-band analysis of speaker variability in linear-prediction cepstra measured from 6 native, male speakers' productions of the 5 Japanese vowels under microphone and mobile conditions. Using a band-selective form of the index-weighted cepstral distance, cepstral F-ratios (between-to-within-speaker variances) were obtained for the entire band from 0 to 3.5 kHz, and for 7 contiguous sub-bands of 500 Hz width each. The resulting profiles of cepstral F-ratios highlight the superiority of certain sub-bands over the full-band and their relative immunity to mobile transmission effects. This is particularly evident in the sub-bands corresponding to the second-formant range for high front vowels.

Index Terms: linear-prediction cepstrum coefficients (LPCC), band-selective cepstral distance, formants, mobile phone speech, forensic voice comparison (FVC).

1. Introduction

In FVC research and casework practice, formants and various types of cesptra, such as LPCC (on linear-frequency scale) or MFCC (on mel-frequency scale), are the most commonly used acoustic features. Various studies which compared FVC performance between formants and cepstrum report that cepstral coefficients generally outperform formants (e.g. [1-3]). This is hardly surprising, since formants represent only the location of the spectral peaks in the frequency domain, whereas cepstral coefficients utilise an entire frequency range (defined by the user) and thus capture richer information both on formants and broader spectral effects. The cepstrum also has the advantage of being extracted automatically. By contrast, automatic formant extraction is known to be highly unreliable, as demonstrated in [4], and usually formants are extracted in a human supervised system, in which the measurer applies manual corrections. This leads to two problems: introducing measurer-dependent variability to the data (see [6-8] for studies and discussions on the difficulties of formant extractions), and being extremely resource intensive.

While they are difficult to extract, formants have 2 major advantages over the cepstrum: robustness and interpretability. Formants are generally understood as more robust than the cepstrum in resisting conditions commonly found in FVC casework, such as channel mismatch and poor audio quality. Formant frequencies also generally correspond to certain articulatory gestures in speech production, and are therefore easier to communicate to the court for what the acoustic features might mean. In FVC casework, experts must communicate the analysis process and evaluation outcomes to non-experts. No matter how excellent the quality of analysis, it is not serving its purpose if experts cannot present their evidence in a way that assists the court in reaching a correct decision. As cepstral features are mathematically derived and hard to link to actual speech production, making it accessible to laypeople is more challenging.

2. Sub-band cepstral approach

As a potentially effective alternative to existing options, this study presents a preliminary investigation of the potential of a sub-band cepstral approach.

The use of the sub-band cepstrum derived directly from the full-band cepstrum was first introduced in [5]. Its potential in speaker classification was then tested in [6] with 6 male Australian English speakers with the word 'hello'. Comparison of F-ratios revealed that the sub-band cepstrum outperforms formants. The same sub-band idea was further tested in an investigation [10] of Japanese vowels spoken by 297 male speakers of Japanese. The results demonstrated a strong similarity between formant-based F-ratios and the cepstrally-based F-ratios obtained in sub-bands spanning the same formant ranges, thus suggesting that the sub-band cepstrum is relatable to acoustic-articulatory gestures.

One of the important characteristics of the sub-band cepstrum is that it allows us to exclude specific parts of a signal that are irrelevant to given speech sounds. Thus it is potentially more robust against poor quality recordings than the full-band cepstrum is. Indeed, using an entire range of spectral slopes, including noise or telephone transmission artefacts, could obscure characteristics of the speech signal.

This characteristic afforded by the sub-band cepstrum becomes particularly attractive to FVC practice when we consider the situation involving mobile telephone speech. While formants are generally considered to be more robust against telephone transmission, they too are not immune from such influences as demonstrated by studies such as [4, 7]. The signal degradation caused by wireless transmission is an obvious problem, but the dynamic nature of the codec can add to challenges. Mobile networks transmit signals packet by packet and, consequently, the conditions of the network, such as congestion, continuously alter the characteristics of the signals received (c.f. [8, 9]). This implies that the channel compensation techniques that rely on the channel characteristics remaining constant may have some limitations.

In this study, therefore, we compare sub-band cepstral variability under two different recording conditions: direct microphone and mobile-to-landline telephone transmission.

We selected 6 male speakers from a large-scale bone-conducted speech database produced by the National Research Institute of Police Science (henceforth NRIPS database) [10], uttering 5 Japanese vowels in 4 different (C)V contexts.

3. Speech material & cepstral analysis

3.1. Database and speakers

Our data were extracted from the NRIPS database [10]. It consists of 332 female and 339 male speakers (671 speakers in total), of which 639 were recorded on two separate occasions, 2 to 3 months apart. All utterances were readout speech, consisting of single syllables, words and selected sentences. The participants range in age from 18 to 76 years. The metadata provide information on which areas of Japan (or overseas in some cases) they have resided, as well as their height, weight, and their health conditions on the day of recording. For this preliminary study we retained 6 speakers with different F0 levels (see Table 1) from the dialectally-homogeneous pool of 52 male speakers who have never lived outside of Tokyo. The mean and standard deviation for the 6 speakers' F0s are 138.9 Hz and 24.8 Hz respectively. This mean appears rather high compared to what is normally considered as a standard male F0, even though the utterances were readout speech.

Table 1. *Speakers' IDs, ages and mean F0s.*

Speaker ID	*M144*	*M072*	*M042*	*M289*	*M295*	*M305*
Age	37	23	20	65	65	71
F0 (Hz)	108.9	109.3	136.6	142	186.7	189.1

The F0s for M295 and M305 are much higher than would be expected for adult males. As shown in Fig.1, one major spectral effect is that the harmonics become sparser with high F0s, which would make accurate formant detection more difficult for these two speakers. It has indeed been demonstrated in [11] that F0 correlates with the range of formant estimation errors. Here we look into the question of whether the sub-band cepstrum can help overcome this difficulty.

3.2. Target utterances

Since this is a preliminary study, we limited our investigation to (C)V sequences produced by the 6 speakers listed above. The *target vowels* are the 5 Japanese vowels /a/, /i/, /ɯ/, /e/, and /o/, in 4 different phonetic contexts: no consonant, /k/+V, /s/+V, /p/+V, thus covering different places (null, velar, alveolar, bilabial) and gestures (plosive and fricative) for the initial-consonant articulation.

3.3. Linear-prediction (LP) cepstrum extraction

First, the vowel nuclei were delimited using the Praat software by interactive auditory-visual inspection of waveforms and spectrograms. The vowel-delimited waveforms were next down-sampled from 44.1 to 11.025 kHz, and subjected to short-term LP analysis (autocorrelation method) with the following conditions: LP-order 14, frame length 25 ms and frame advance 5 ms, Hamming windowing, and a pre-emphasis factor of 0.98. Standard LP-cepstrum coefficients (LPCCs) were finally obtained recursively from the LP-autoregressive coefficients for each frame of a vowel nucleus and then averaged over all frames within that nucleus. Note that the LP procedure described above effectively yielded LPCCs that represent smoothed spectral information over the full range from 0 to half of 11.025 kHz. Hence we call these *full-band* LPCCs.

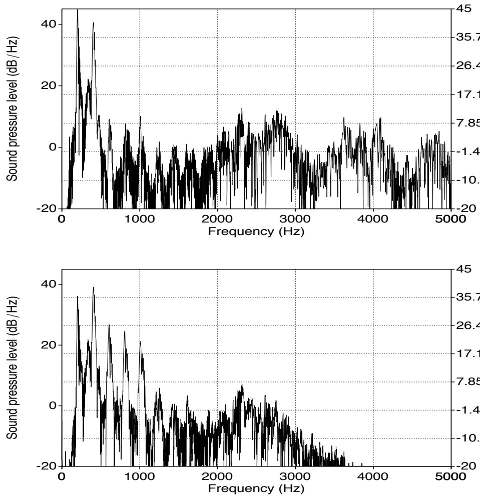


Figure 1. *Long-term spectra of speaker M305's /i/ recorded through microphone (top panel) and mobile network (bottom panel).*

4. Cepstral F-ratio procedures

The ratio of between- to within-speaker variances, here referred to in short as the (speaker) F-ratio, was implemented in the cepstral domain with the aims: a) to uncover the sub-band(s) of a vowel's smooth spectrum that provide the greatest amount of speaker-specific information; and b) to determine the extent of mobile transmission effects in such bands.

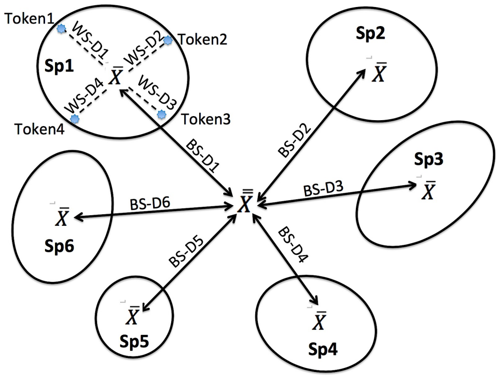


Figure 2. *Graphical procedure for calculating within- (WS) and between-speaker (BS) cepstral variances expressed as Euclidean distances (D).*

4.1. Full-band procedure

Figure 2 helps to visualise the basic procedure for calculating the between- and within-speaker variances, which are mathematically described by Eqs. (1) and (2), respectively:

$$\sigma^2_{between} = \frac{\sum_{i=1}^{N}(\overline{X_\iota} - \bar{\bar{X}})^2}{N-1} \qquad (1)$$

$$\sigma^2_{within} = \frac{\sum_{i=1}^{N}\sum_{j=1}^{n_i}(X_{ij} - \overline{X_\iota})^2}{\sum_{i=1}^{N} n_i - N} \qquad (2)$$

where n_i is the number of tokens (4) per speaker, N is the total number of speakers (6), X_{ij} is a vector of LPCC coefficients of order 14, and where:

$$\bar{X}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} X_{ij} \text{ is the mean for the } i^{th} \text{ speaker}$$

$$\bar{\bar{X}} = \frac{1}{N}\sum_{i=1}^{N} \bar{X}_i \text{ is the grand mean}$$

Note that the numerators in Eqs. (1) and (2) are Euclidean distances between pairs of cepstral vectors. *Full-band* F-ratios were obtained with the full-band LPCCs in these distances.

4.2. Sub-band procedure

Sub-band variances involve the band-selective formulation derived in [5] for weighting the distances in Eqs (1) and (2) according to the lower and upper limits of the sub-band selected. Note that the formulation affords the flexibility of re-using the full-band LPCCs without needing to return to the acoustic analysis stage in order to generate new LPCCs every time a new sub-band is of interest. It also affords the possibility of selecting any sub-band of any width within the full band.

For this preliminary study, contiguous sub-bands of equal width (0.5 kHz) were chosen to span the range from 0 to 3.5 kHz, a choice motivated by the observation (see bottom panel of Fig. 1) that the mobile network tended to obliterate spectral information above 3.5 kHz.

5. Results and discussion

Fig. 3 presents the F-ratios for both microphone and mobile recordings in the 7 sub-bands. Since the comparison of the spectra presented above revealed a rapid decay of the signal in the frequency range above approximately 3.3 kHz for the mobile data, we limit our analysis range to 0-3.5 kHz on this occasion. We pooled the results for the front vowels /i/ and /e/ (top panel), and the back vowels /a/, /o/, and /ɯ/ (bottom panel). For comparison, we also plotted the F-ratio derived from the cepstral differences of full-band LPCCs (horizontal lines). Because the F-ratio is essentially the ratio of between- and within-speaker variabilities, a high F-ratio can be said to indicate the potential for stronger speaker discrimination.

Overall, we can see that the 2 channels behave rather similarly below the 2 kHz region. The wide discrepancy between the channels above this region is attributed to mobile transmission. Indeed, as presented earlier in Fig. 1, the spectral energy was heavily damped from around 2.2-2.5 kHz in the mobile recordings. The top panel (front vowels) clearly indicates that the F2 region (1.5-2 kHz) has strong speaker discrimination potential under both conditions, although it should be noted that microphone recordings produced higher F-ratios in general. For back vowels in the bottom panel, the F-ratio within the F2 region is very weak, but the F3 range (2-3 kHz) for the microphone recordings shows a high F-ratio.

The correspondence between the F-ratios presented here and previous studies is remarkable. For instance, a study on Japanese vowels [12] reported that F2 of /i/ and F2 and F3 of /e/ were reported to have substantially stronger F-ratios compared to other vowel/formant combinations. For back vowels, F-ratios were relatively low in general, but with /o/ F3 was reported to be most useful for speaker classification. The study on the formant F-ratios of Australian English [13] also reported the same. A large scale study on Japanese vowels with sub-band cepstrum reports the same outcome [14].

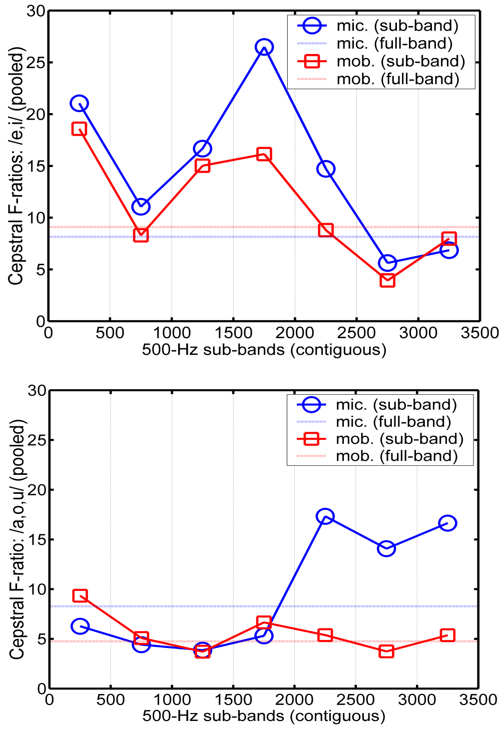


Figure 3. *Full-band and sub-band cepstral F-ratios for front (top panel) and back (bottom panel) vowels.*

Another observation is that the selected sub-bands perform much better than the full-band cepstrum. One curious finding is that, for the front vowels, the full-band cepstrum of the mobile recordings slightly outperformed that of the microphone recordings, contrary both to our expectation and to the results for the back vowels. However, the observed difference between the two channels seems too small for the front vowels to ascertain its significance. Further investigation with more speakers is necessary.

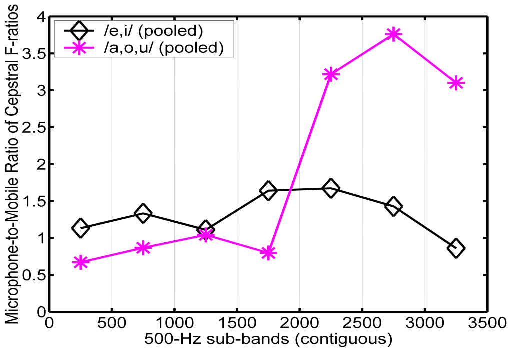


Figure 4. *Microphone-to-mobile ratios of the sub-band cepstral F-ratios shown in Figure 3.*

Fig. 4 presents the microphone-to-mobile ratios of the F-ratios presented in the previous figure. It is quite clear that speaker discrimination would be severely degraded for the mobile data in the higher sub-bands. Mobile phone transmission appears not to affect the bands below 2 kHz by very much. In the region above this, however, there is a

Table of Contents for this manuscript

marked difference between front and back vowels. As seen in Fig. 3, the front vowels did not produce high F-ratios in this region even with the microphone recordings, so the mobile transmission had very little effect on the F-ratio performance. Back vowels, on the other hand, had very strong F-ratios in the F3 region for microphone recording. Thus the impact of mobile transmission becomes more evident in this region.

To sum up, the sub-band has a few interesting properties. First, its behaviour corresponds to the known behaviour of the formants. Selected carefully, the sub-band could produce more powerful features for speaker differentiation than those based on the full-band cepstrum. Further, it behaves very similarly across the two channels as long as its sub-bands are within the transmission's passband. Also the microphone-to-mobile degradation in the F-ratios was not very high. This, together with the findings in previous studies, suggests that the sub-band cepstral approach could provide us with a way to overcome the current challenges associated with the use of either the formants or of the full-band cepstrum.

We also suspect that the sub-band cepstrum could be more robust against added processing such as sound enhancement or channel mismatch. Through past casework, the authors have experienced that acoustic pre-processing of sound recordings can shift formant frequencies, even though the voice quality seems unchanged auditorily. We speculate the nature of LPC formant detection to be the cause, as it relies on identifying peaks in the spectral slopes. The exact location of the peaks on the frequency axis could be susceptible to added effects such as the boosting of particular frequency ranges. By contrast, the overall smooth shapes captured by the cepstrum are much less sensitive to these effects.

6. Conclusions and ways forward

The data used in this study is limited and, therefore, it is not yet possible to draw very definite conclusions. However, it is our view that the promise of the sub-band cepstral approach for FVC was sufficiently demonstrated in this preliminary study. We thus plan to extend this work in several ways. Needless to say, we need testing with a much larger speaker set. We must also investigate the cross-sessional and cross-channel F-ratios to increase forensic realism. Finally, the sub-band selection needs to be evaluated more extensively. This study used 500 Hz equidistant bands, but we hypothesize that changing the sub-band ranges to those spanning each vowel's expected formant ranges will allow us to focus more sharply on the acoustic information relevant to forensic settings.

An international survey on FVC practice published in 2011 reports that 97% of the survey participants use formants [15], making formants one of the most favoured acoustic features in FVC casework. However, as discussed, the difficulty of reliable formant extraction and its time consuming nature justify the search for a more effective approach. These issues are even more accentuated under the current paradigm of FVC, where a substantially sized background population is an essential requirement for reliable assessments. In this light, we believe that the sub-band cepstral approach could be an alternative acoustic feature with the potential to significantly improve the reliability and validity of the FVC process, not least through the more efficient analysis of more comprehensive reference data.

7. Acknowledgements

We thank Dr Phil Rose for inspiring discussions on the sub-band approach and its potential for more robust FVC. We also thank Dr David Broad for helpful suggestions related to content and style.

8. References

[1] 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.

[2] 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.

[3] 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.

[4] 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.

[5] 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.

[6] P. Rose and F. Clermont, "A comparison of two acoustic methods for forensic speaker discrimination," *Acoustics Australia,* vol. 29, pp. 31-35, 2001.

[7] 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.

[8] E. Alzqhoul, B. B. Nair, and B. J. Guillemin, "Speech handling mechanisms of mobile phone networks and their potential impact on forensic voice analysis," in *The 14th Australasian International Conference on Speech Science & Technology*, Sydney, 2012, pp. 29-32.

[9] B. J. Guillemin and C. Watson, "Impact of the GSM Mobile Phone Network on the Speech Signal – Some Preliminary Findings," *The International Journal of Speech, Language and the Law,* vol. 15, pp. 193-218, 2008.

[10] 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.

[11] 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.

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

[13] P. Mokhtari and F. Clermont, "A Methodology for investigating vowel-speaker interactions in the acoustic-phonetic domain," in *The 6th Australian International Conference on Speech Science and Technology*, Adelaide, 1996, pp. 127-132.

[14] 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.

[15] E. Gold and P. French, "International practices in forensic speaker comparison," *International Journal of Speech, Language & the Law,* vol. 18, pp. 293-307, 2011.

Visualisation tools to analyse phonetic confusions for speech perception tests

Chung Ting Justine Hui[1], *Catherine Watson*[2], *Takayuki Arai*[1]

[1] Sophia University, Japan
[2] University of Auckland, New Zealand

justinehui@eagle.sophia.ac.jp, c.watson@auckland.ac.nz, arai@sophia.ac.jp

Abstract

To enhance the intelligibility of synthetic voices, especially for the hard-of-hearing, these voices need to be evaluated in a phonetically systematic way. This enables identification of problematic sounds and confusions that come with them, thus allowing suitable adjustments to the voice. Firstly, modified forms of confusion matrices are introduced to examine and compare phonetic confusions. This is followed by the consonant confusion cloud plots, developed to visualise confusion data by focusing on individual consonants. These tools enable us to identify the distribution and relationship of confusion with the target phone in terms of manner, place and voicing in one glance.

Index Terms: speech perception, phonetic confusion, speech synthesis, confusion matrix, consonant confusion cloud

1. Introduction

Synthetic voices need constant evaluation to ensure they are being perceived accurately. This gives rise to different types of perception tests, especially designed to evaluate the intelligibility of these voices. One of the more utilised tests is the semantically unpredictable sentences, applied in the Blizzard challenges, one of the international recognised benchmark test that invites researchers to present synthetic voices annually and be judged by a public perception survey [1, 2, 3, 4]. While semantically unpredictable sentence perception tests are able to tell us the intelligibility of the voices on an utterance level, it lacks the granularity in focusing on individual phones to allow for possible enhancement on the voices on a phonemic level.

Wolter, on her studies to access synthetic voices on healthcare robots, took a different approach and evaluated the voices using unfamiliar medication names [5, 6]. From the answers produced by the participants, she then analysed the phonetic error in their perceived answers. Taking inspiration from this, a similar perception test has been carried out previously using pseudo medicinal names to test phonemic intelligibility of synthetic voices collected from 160 participants [7].

Part of the test examined 38 English consonants and consonant clusters in non word-final positions as shown in Table 1. Having this granularity to examine how the consonants produced by the synthetic voices are being perceived, we can then make suitable enhancements to the voices to increase their intelligibility. This gives rise to the need for tools for examining the errors the participants are making as well as the relationship of these errors with the target phone.

p pl pɹ t tɹ k kl kɹ b bl bɹ d dɹ g gl gɹ
f fl fɹ θ θɹ s ʃ sm sn st stɹ h v z ʧʤm n w j l ɹ

Table 1: *List of non-word final consonants*

For a start, the confusion matrices introduced by Miller and Nicely in 1955 act as an adequate tool to examine all the consonants and the confusion in the participants' perception [8]. However, when we add in multiple variables, for example, to compare between two different synthetic voices, or two groups of participants, the traditional confusion matrix may not be able to handle the multitudinal nature of the data.

This paper describes such visualisation tools to enable researchers to analyse the perception results; from modifying the traditional confusion matrices to applying the concept of tag clouds to display the phonetic data. We will be using data obtained from a previous perception study [7] and discuss briefly observations made from the plots to highlight the functions of these tools.

2. Background data

While this paper does not focus on how the data was collected, this section will briefly introduce the data to be analysed by the visualisation tools in following sections [7].

In a previous study, perception data were collected from 160 participants to investigate the impact of hearing loss on the intelligibility perception of English consonants. This paper uses the data from one of the intelligibility tests where participants were asked to spell out the pseudo medicinal words as they hear them. 38 consonant and consonant clusters were tested in a series of sentences in the form of ”At 9:30am, please take your [medication]”, where medication is taken from a list of made-up medicine names consisting of the consonants listed in Table 1.

The participants were separated into groups according to their attributes such as first language English speakers and second language English speakers, hearing impaired (HI) and normal hearing (NH) and participants who are above the age of 60 and under the age of 60. This paper uses the data collected from 160 participants, dividing the participants into those who experience hearing loss (89) and those who do not (71).

Three voices were used to pronounced the stimulus sentences, two synthetic voices and one natural voices. For the purpose of this study, data from all three voices were combined.

2.1. Correct identification analysis

Before we launch into discussing the confusion participants make when they could not identify the consonants correctly, let us have a look at the correct identification rate for each of the consonants in Table 1. Figure 1 shows the list of consonants and their respective accuracy rate in terms of the hearing impaired and the normal hearing participants. We can see that from this simple text plot, how well each consonant is being recognised, and those they are more difficult to understand. Unsurprising,

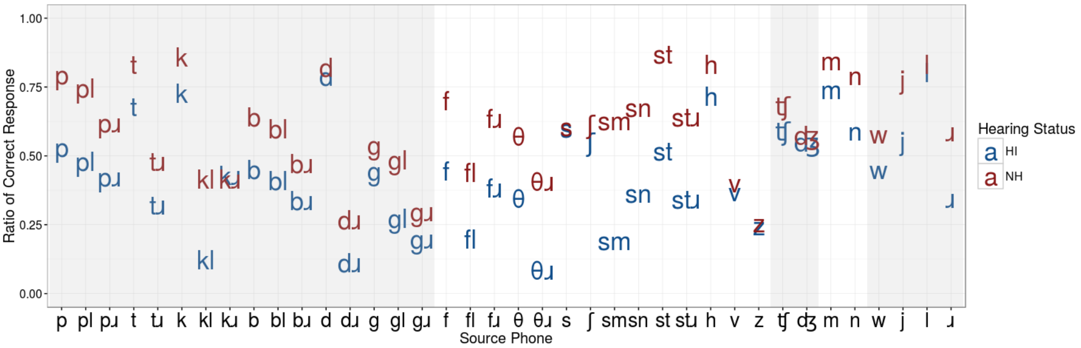


Figure 1: *Consonant identification plot for non word-final consonants*

hearing impaired participants could identify all the consonants less correctly than their normal hearing counterparts. Consonant clusters seem to especially cause distress, where less than 20% of hearing impaired participants could not identify /kl/, /dɹ/ and /θɹ/. On the other hand there are consonants such as /kɹ, /v/ and /z/ where they are similarly difficult for both groups.

3. Visualisation tools to aid analysis

3.1. Aim

While Figure 1 can offer us an insight into what consonants are difficult to perceive for the hearing impaired participants compared to their normal hearing counterparts, we need to be able to identify what the sound they confuse the target phone with and the phonetic relationship between their mistakes and the source phones to ultimately be able to enhance the synthetic voices. For example, if we find that voiceless stops are being mistaken for voiced stops, we might be able to manipulate the voice onset time or the length of the following vowel to make the consonant easier to identify.

This brings us to the need for a set of tools for developers to evaluate the synthetic voice in order to enhance speech signal, especially for the hearing impaired to understand. The following sections will describe these visualisation tools that have been used to display the errors the participants have made in a more systematic way from a phonetic perceptive. Due to restriction on space, the consonant clusters are left out in the plots. For details on the consonant clusters, please see [7].

All figures are generated using R [9], and the visualisation tools were developed from scratch with the use of packages wordcloud [10] and ggplot2 [11].

3.2. Confusion matrices

Modified from the traditional confusion matrices as seen in Miller (1955), Pollack (1960) and Singh (1966) to name a few [8, 12, 13], the confusion matrices in this paper are overlaid with a heat plot where the opaqueness of the tiles represent the number in ratio of participants perceiving a particular phone and the colours represent the two hearing status groups.

The confusion matrices combines the correct identification rate of the consonants and the consonant choices for any particular phones into a bird's eye view of the phoneme error analysis for the intelligibility tests.

Figure 2 and Figure 3 allow us to examine all the consonants identified and confused by the participants, separating the data hearing impaired (HI) and normal hearing (NH) groups. The x-axis of the confusion matrix represents the targeted phones and the y-axis lists out the answers given by the participants. Two arbitrary symbols were chosen to represent vowels, a 'v' enclosed in a circle, and invalid answers, a dash across a circle. Due to the size of the tiles, the ratio has been rounded to 1 decimal point. This means that a '0' tile indicates less than 5% of answers were identified as the specific phone, as opposed to an empty tile, where there were no answer.

Diagonally we can see the a relatively strong response where the participants are identifying the source phone correctly. By combining the numbers and the opaqueness of the tiles, we are then able to identify the problematic phones, and their confusion from the columns.

We can see that while there are some strong candidates identified 80% and above correctly by the hearing impaired participants such as /d/ and /l/, most of the time the diagonal tiles are faint in their colour, indicating a low correct identification rate. We can also observe some clustering of confusions around the sonarants, showing that the confusions for sonorants occur within amongst themselves. On the other hand, the diagonal tiles for the normal hearing results are much more opaque than the hearing impaired results, being in consistent with out results from Figure 1. However, for both groups, it seems that /z/ is a difficult sound to identify in this particular test, with more than 50% of the answers being confused with /s/ instead.

3.3. Combined confusion matrices

Now, what happens if we want to compare the two groups together to show how the differences and similarities in the mistakes the participants make? In the combined confusion matrices, we make use of the colour hue and again opaqueness of the tiles to show both groups on the same confusion matrix, as shown in Figure 4, where again red represents the normal hearing results and blue the hearing impaired results.

A tile that has a purple hue describes both groups having similar levels of difficulty in identifying the phone, a red hue indicates the normal hearing has a stronger presence and a blue hue for the hearing impaired group. From Figure 4, we can approximate the normal hearing having a larger varieties in their confusion, where the tiles away from the expected diagonal line display more 'red' than 'blue'. On the other hand, the hearing impaired group has a higher ratio of invalid answers, shown by the more 'blue' hue along the last row of the matrix.

However, while we can observe all the tested phones in one glance, it is difficult to observe the relationship and distribution between each individual consonant and their confusions. This brings us to the next type of plots, the consonant confusion clouds, where we can focus on individual consonant at a time.

Table of Contents for this manuscript

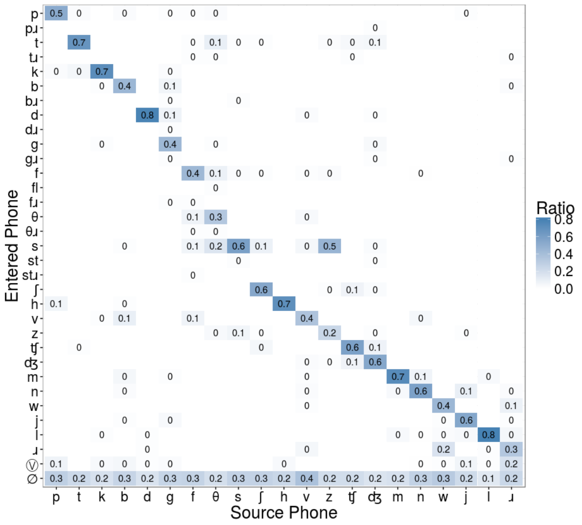


Figure 2: *Confusion Matrix for hearing impaired participants*

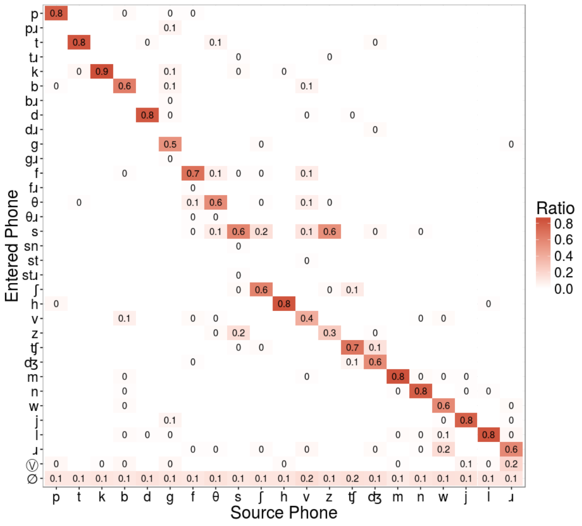


Figure 3: *Confusion Matrix for normal hearing participants*

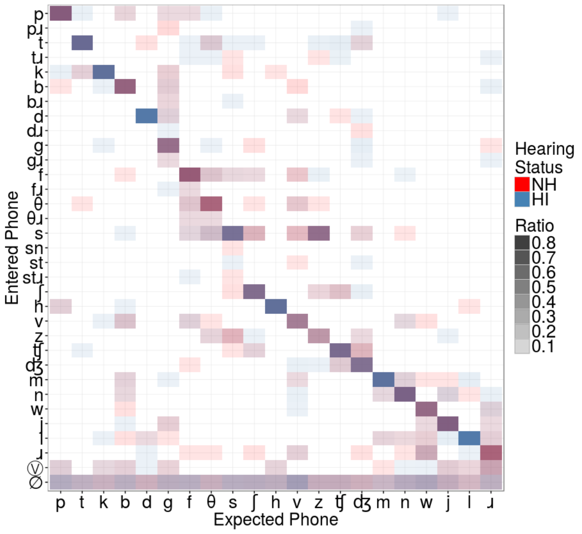


Figure 4: *Confusion Matrix displaying results from both hearing impaired and normal hearing participants*

3.4. Consonant confusion clouds

The consonant cloud plots take inspiration from tag clouds, which are typically used to depict metadata or tags in websites where words are mapped onto a cloud of words with their importance or occurring frequency signified by font size [14]. In the same way, the frequency in which the sounds are identified by the participants governs the font size of phones in the consonant cloud plots.

Unlike the consonant identification plots or the confusion matrices, the cloud plots are not able to indicate the absolute value of how accurately a phoneme is being perceived. Instead, it gives a graphical representation of all the sounds heard mistakenly in an arbitrary consonant grid where the x-axis represents the location of articulation and the y-axis represents the manner of articulation and voiced and voiceless pairs are presented side by side as shown in Figure 5. By having an arbitrary grid reliant on the three characteristics of a consonant: the manner, location and voicing, we may be able to identify a pattern in which the sounds are being confused with. The code to generate these cloud plots has been written from scratch with the help of the R package wordcloud [11].

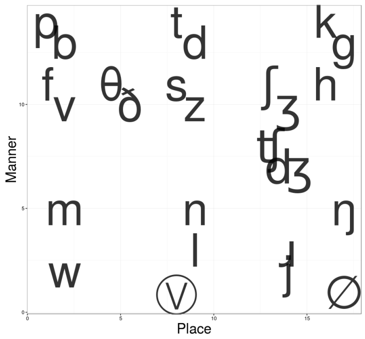


Figure 5: *How the consonants are placed according to their manner and place of articulation on the cloud plots*

Using colours of the font, the cloud plots are able to compare multiple groups, such as how two or more synthetic voices compared to each other, or how different groups of participants compare in their perception for individual consonants.

The font size is determined by an arbitrary inverse square-root relationship with the number of participant to make the less frequent consonants visible. Due to the nature of clutter of the cloud plots, the current set up makes it difficult to include the consonant clusters. International Phonetic Alphabet symbols are used, and again the two symbols used for vowels and invalid answers as in the confusion matrices are applied here.

Let us take a look at Figure 6, showing the answers from the hearing impaired and the normal hearing group perceived for stop consonants. We can see that for /t/, /d/ and /k/, the majority for both groups could identify the correct consonant fairly accurately by the large font of the letters, with little confusions. When we compare the velar stop pair, /g/ seems to trigger many more variation of other phones than /k/ for both groups of participants. While both groups seem to confuse the /g/ sound with /b/ similarly, there is a tendency for the normal hearing to confuse /g/ with /k/, and the hearing impaired participants to confuse /d/ with /g/ than their counterparts.

For all the stops apart from /t/, we can observe the hearing impaired participants producing more invalid answers than their normal hearing counterparts from their sizes of the ”invalid” symbol, suggesting that the hearing impaired had trouble

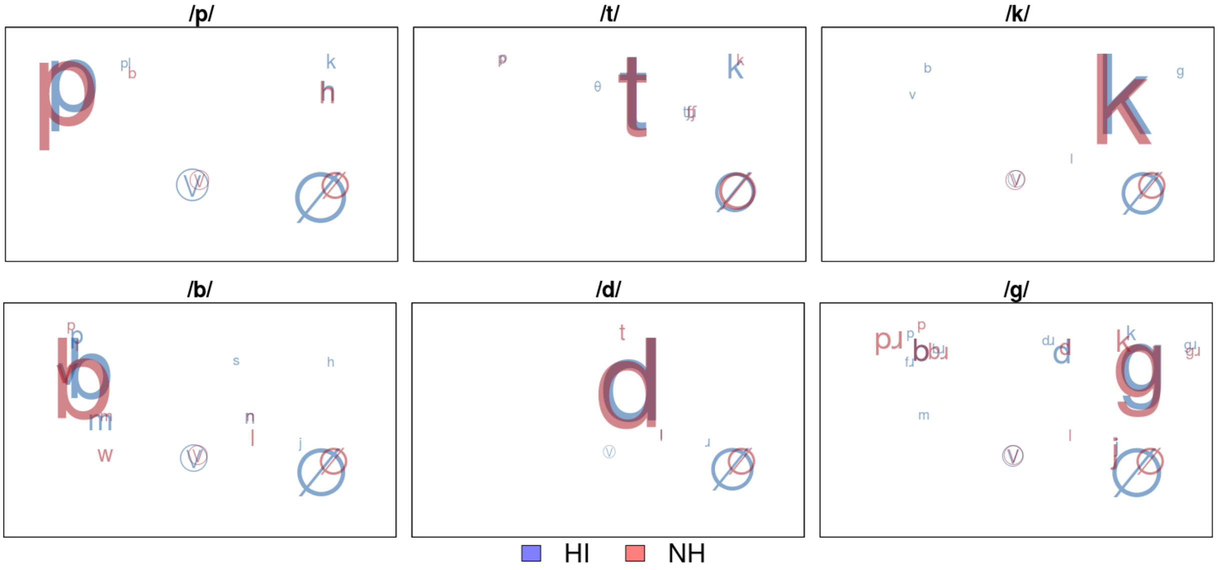


Figure 6: *Consonant identification cloud plots of stops*

hearing the phone at all to take a guess at the question.

We can see from these figures that most confusions only differ in one aspect from the target consonant, that is, the mistake occurs either in identifying the consonant's manner, place of articulation or voicing. For example, we can see in /b/ that the confusions tend to be manner of articulation, with mistaken sounds being /m/, /w/, /v/. On the other hand, /g/ has more places of articulation confusion, such as /d/ and /b/, and the insertion of /ɹ/. Both consonants also have a small percentage of voicing confusion with the consonant being mistaken as their unvoiced counterparts, /p/ and /k/.

These cloud plots allow us to be able to analyse the confusion data with a greater focus than the confusion matrices, helping us to locate the problematic sounds and their confusions leading to possible enhancement for the voices.

4. Conclusions and other applications

Using the data gathered from a previous study, we are able to locate the problematic phones and mistakes participants are making using the confusion matrices and the consonant confusion clouds. These visualisation tools give us a greater level of granularity to examine the phones that the participants may have trouble deciphering. On top of that, the cloud plots focus on each individual phones, allowing us to observe the confusion using phonetic knowledge. Preliminary observation shows that most errors only differ in one aspect of the target consonant, be it manner, place or voicing.

Finally, while these tools were designed to evaluate synthetic voices, the use of these tools does not need to be limited to synthetic voices and can be applied to other perception tests when examining consonant confusions.

5. Acknowledgements

We would like to thank our participants and Triton Hearing for their support in recruitment for this study.

6. References

[1] A. W. Black and K. Tokuda, "The Blizzard Challenge — 2005: Evaluating corpus-based speech synthesis on common datasets," *Interspeech 2005: 6th Annual Conference of the International Speech Communication Association*, pp. 77–80, 2005.

[2] K. Tokuda, Y. Nankaku, T. Toda, H. Zen, J. Yamagishi, and K. Oura, "Speech synthesis based on hidden Markov models," *Proceedings of the IEEE*, vol. 101, no. 5, pp. 1234–1252, 2013.

[3] Z.-h. Ling, Y.-j. Wu, Y.-p. Wang, L. Qin, and R.-h. Wang, "USTC System for Blizzard Challenge 2006 an Improved HMM-based Speech Synthesis Method," *Blizzard Challenge Workshop*, 2006.

[4] H. Zen, T. Toda, and M. Nakamura, "Details of the Nitech HMM-Based Speech Synthesis System for the Blizzard Challenge 2005," *IEICE transactions on information and systems*, vol. 90, no. 1, pp. 325–333, 2007.

[5] M. Wolters, P. Campbell, C. Deplacido, A. Liddell, D. Owens, and A. Division, "Making Speech Synthesis More Accessible to Older People," in *Sixth ISCA Workshop on Speech Synthesis*, 2007.

[6] M. Wolters, P. Campbell, C. DePlacido, A. Liddell, and D. Owens, "The Effect of Hearing Loss on the Intelligibility of Synthetic Speech," *International Congress of Phonetic Sciences*, vol. 16, no. August, pp. 673–676, 2007.

[7] C. T. J. Hui, "Development and Implementation of a Perception Toolkit to Evaluate the Impact of Synthetic Speech on the Hearing Impaired," Master Thesis, University of Auckland, 2016.

[8] G. A. Miller and P. E. Nicely, "An Analysis of Perceptual Confusions Among Some English Consonants," *Journal of the Acoustical Society of America*, vol. 27, no. 3, pp. 338–352, 1955.

[9] R Core Team, "R: A language and environment for statistical computer," 2015. [Online]. Available: https://www.r-project.org/

[10] I. Fellows, "wordcloud: Word Clouds," 2014. [Online]. Available: http://cran.r-project.org/package=wordcloud

[11] H. Wickham, *ggplot2: elegant graphics for data analysis*. Springer New York, 2009. [Online]. Available: http://had.co.nz/ggplot2/book

[12] I. Pollack and L. Decker, "Consonant Confusions and the Constant Ratio Rule," *Language and Speech*, vol. 3, pp. 1–6, 1960.

[13] S. Singh and J. W. Black, "Study of Twenty-Six Intervocalic Consonants as Spoken and Recognized by Four Language Groups," *Journal of the Acoustical Society of America*, vol. 39, no. 2, pp. 372–387, 1966.

[14] M. Halvey and M. T. Keane, "An Assessment of Tag Presentation Techniques," 2007. [Online]. Available: http://www2007.org/htmlposters/poster988/

Japanese Vowel Deletion Occurs in Words in Citation Form

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

[1]University of Melbourne, Australia
[2]MARCS Institute for Brain, Behaviour and Development,
Western Sydney University, Australia

akilpatrick@student.unimelb.edu.au, rikkelou@gmail.com, bjbaker@unimelb.edu.au

Abstract

Japanese vowels have allophonic reduced variants, including shortened, devoiced and deleted instances. This kind of linguistic behaviour is commonly associated with rapid or casual speech. However, the present report demonstrates that vowel deletion also occurs when Japanese speakers produce Japanese words in citation form, and not just in rapid or casual speech. We propose that deletion is more likely to occur in high-frequency lexical items, specifically in three regularly occurring suffixes.

Index Terms: Japanese, vowel deletion, elision

1. Introduction

Cross-linguistically, segmental reduction is typically associated with fast speech. In Japanese, however, vowel reduction—shortening, devoicing, and deletion—is said to occur regularly for the high vowels, /i/ and /u/, even at a normal speaking rate [1]. Shortening, devoicing, and deletion behaviours are generally regarded as features of a single phenomenon, referred to as 'vowel devoicing' or 'reduction'. Reduced vowels are prominent in many Japanese dialects though they characterise the dialects of Kanto and Kyushu in particular [2]. The existing literature suggests that devoicing is more likely to occur when a vowel is preceded by a stop while deletion is more likely to occur when a vowel is preceded by a fricative [3]. [1] argues that it is impossible to determine whether certain vowels are devoiced or deleted, particularly when they occur after fricatives, even by examining the acoustic signal.

In the following, we present a study of the production of Japanese words in citation form and argue that deletion is a feature of citation form and is more likely to be elicited by specific words that occur in high frequency in Japanese discourse.

2. Background

Cross-linguistically, vowel production is typically characterised by regular vocal fold vibration (voicing). However, Japanese exhibits a range of devoicing patterns that challenge this observation. Japanese high vowels are regularly devoiced when they are both preceded and followed by unvoiced consonants or when they appear the end of a word and are preceded by an unvoiced consonant. Japanese vowels are also said to undergo deletion in certain environments, although there is some contention regarding this. Those reports [3] that argue that deletion is occurring, suggest that vowels are deleted when they follow fricatives and devoiced when they follow stops. Other reports [2] treat deletion as an extreme variant of devoicing, suggesting that reduction is a continuum, at one end of which, vowels are voiced and maintain regular duration and deleted at the other, with devoicing occurring at a midpoint.

It is unlikely that Japanese vowel reduction is a recent development. From as early as the 1600s, non-Japanese scholars have been remarking that Japanese high vowels are sometimes inaudible [2]. However, devoicing patterns in Japanese vary across regional dialects: Systematic reduction is a feature of the Kanto dialect but far less frequent in the Kansai dialect [4]. Indeed, cities in East Japan have higher devoicing rates, 68% in Nagoya, 56% in Tokyo, and 56% in Sendai, than cities in the Kansai and Shikoku regions of central Japan, 32% in Osaka, 28% in Okayama, and 18% in Kochi, however, further west, these rates increase to those levels found in Eastern cities, 57% in Naha and 53% in Kumamoto [5].

Vowel reduction behaviour, including devoicing, primarily occurs with the high vowels, /i/ and /u/ [6]. Reduction may, however, also occur with other vowels but this is irregular, arising only in specific lexical items. A frequently cited example is the devoicing of the first /o/ in the Japanese word kokoro [*heart*], as is the first /a/ in katana [*sword*] [3]. Irrespective of the vowel, reduction occurs in two environments; when a high vowel is both preceded and followed by a voiceless consonant, and when a high vowel occurs word finally and is preceded by a voiceless consonant.

Japanese verbs largely follow predictable conjugation rules, though there are a small number of exceptions. The polite, present tense suffix—ます/masu/—conjugates by deleting the word-final mora (or final two mora in exceptions) whereas imperative and past tense suffixes—て/te/ and た/ta/ respectively—conjugate differently, depending on the word-final mora. For instance, verbs ending in す/su/ or する/suru/ conjugate by deleting the word-final mora and adding して/ʃite/ or した/ʃita/ but conjugate differently for verbs ending in alternative mora.

3. Method

3.1. Participants

Twelve native Japanese speakers (10 female; 2 male) living in Melbourne, Australia, were recruited for this study. Seven of the participants were expatriates, having lived in Australia for a minimum of five years. Expatriate participant ages ranged from 27-42. The remaining five participants were international students who had lived in Australia for less than a year and were studying English as a Foreign Language at La Trobe University. International Student participants ranged in age from 18-20.

Participants were recruited by word of mouth. Four of the participants indicated that they were native to districts of Japan where the regional dialects are characterised by lack of vowel reduction (see above). These included the Kansai and Shikoku regions. The remaining eight participants indicated that they were native to areas where vowel reduction is a common feature of the regional dialects, these included Tokyo, Yamagata, Nagoya and Kyushu.

3.2. Recordings

Recordings were conducted in quiet rooms in Melbourne, using a Zoom H4n recording device with a sampling depth of 24kb/sec and a sample rate of 44.1kHz. Participants were asked to read from a pseudorandomised list of thirty words in Hiragana script. Participants received no coaching in how they were to pronounce the list of words, and we provided no feedback on their productions.

The word list contained lexical selected to induce deletion or vowel reduction, on the basis of the existing literature [2], [3]. Indeed, we hypothesised that participants would delete vowels in the following contexts:

- /i/ in the imperative verb suffix in e.g. して /ʃite/ [*IMP*],
- /i/ in the past tense verb suffix in e.g. した /ʃita/ [*PST*],
- /u/ in the polite, present tense verb suffix in e.g. ます /masu/ [*PRS*].

We included a number of words containing each of these suffixes (see Table 1 for the list). Amongst the words, we included the suffix by itself to determine whether there was a difference between the suffix presented in isolation and when attached to a verb. We also included three tokens often reported as examples of low vowels that frequently undergo devoicing, こころ/kokoro/ [*heart*] [1], ほこり/hokori/ [*pride*] [7], and はか/haka/ [*grave*] [8]. Two words, しだ/ʃida/ [*fern*] and しで /ʃide/ [*paper streamer*], were included to test the influence of having a voiced consonant follow the vowel in question. Four words, からす/karasu/ [*crow*], ましʃ/maʃi/ [*better*], すき/suki/ [*like*], and だいすき/daisuki/ [*really like*] were included to test whether deletion would occur in similar but not identical environments to the suffixes listed above. Seven foils were chosen at random from a Japanese dictionary. We included a single instance of word with the post-stop environment: もく /moku/ [*wood*], to test the difference (if any) between fricatives and stops. Speakers produced three repetitions of each of the 30 items, resulting in a total of 90 tokens for analysis per participant, and a total of 1080 tokens included in the present study.

3.3. Predictions

We predict that speakers will delete high vowels when they are preceded by voiceless fricatives and either occur word finally, or followed by voiceless consonants. Furthermore, we predict that deletion will occur more frequently in these environments when they occur in one of the three suffixes, して /ʃite/ [*IMP*], した /ʃita/ [*PST*] and ます /masu/ [*PRS*], and that this is likely due to their high frequency of occurrence.

3.4. Analysis

Elicitations were categorised on the basis of visual inspection of spectrograms. Because there is controversy regarding the existence of deleted vowels in Japanese, we provide the following detailed explanation as to how we categorised participant behaviour. Voiced vowels are characterised by voicing bars—which are indicative of vocal fold vibration—and formant resonance in spectrograms. Unvoiced vowels do not yield voicing bars but do generate formant resonance. Figure 1 shows three instances of the elicitation of the Japanese word, だいすき /daisuki/ [*really like*], produced by one of our participants, in succession. Across all participants, we observed that the /u/ in this particular environment was frequently deleted. In the first iteration, there are clear voice bars; there is considerable energy in the lower frequencies; and a substantial increase in intensity. While this is a shortened, low intensity vowel, the voice bars indicate that the vocal folds are vibrating and hence that the vowel is voiced. We provide the second iteration as an example of a devoiced /u/. Here, we observe a lack of voice bars and a decrease in both low frequency energy and intensity when compared to the first iteration. There is a small but very clear increase in intensity in this devoiced example. In the third example, there is no vowel. There is no transition from the fricative until the sudden reduction of energy which indicates closure for the following stop. The intensity shows a steady weakening with no instances of incline as can be seen in the second example. It should be noted here that the second example in Figure 1 was the only elicitation in which we observed a devoiced vowel following a fricative.

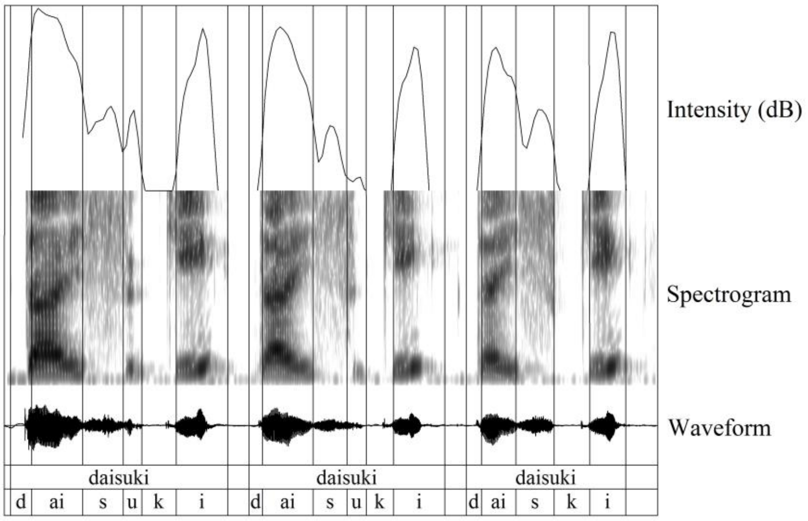


Figure 1: *Voiced, devoiced and deleted examples of the /u/ in the Japanese word, daisuki.*

4. Results

As predicted, participants frequently deleted vowels in the three suffix contexts. This was more frequent when the suffixes were attached to lexical items (*M* = 85%) than when they were uttered as stand-alone words (*M* = 69%). Interestingly, both しゅて /ʃute/ [*nonce*] (19%) and しゅてん /ʃuten/ [*red mark*] (33%) elicited instances of deletion, creating near homophony between して /ʃite/ [*IMP*] (83%) and しゅて /ʃute/ [*nonce*]. As expected, しだ /ʃida/ [*fern*] and しで /ʃide/ [*paper streamer*], which were included to test the influence of having a voiced consonant follow the vowel in question, prompted no instances of deletion. Words with similar but not identical environments to the suffixes did elicit deletion, however this was much more common with medial vowels in すき /suki/ [*like*] (64%) and だいすき /daisuki/ [*really like*] (78%) than in word final vowels in からす /karasu/ [*crow*] (17%) and まし /maʃi/ [*better*] (17%). Those words included as instances where non-high vowels undergo devoicing provided few elicitations of this behaviour: こころ /kokoro/ (6%) [*heart*], ほこり /hokori/ (8%) [*pride*], and はか /haka/ (0%) [*grave*].

There was very little difference between the behaviours of those participants from regions where vowel reduction is common and those from regions where it is reportedly infrequent. Participants from the Kansai and Shikoku regions where reduction is infrequent, showed the same level of deletion as those from regions where reduction is common, both groups displayed deletion in 35% of tokens. While these results do show no effect of regional dialect on citation form, one participant exhibited considerable individual difference which warrants consideration. This participant, who indicated that she spent her childhood in Kanagawa just south of Tokyo, produced deletion in only two instances. This individual was the only participant not to delete the /i/ in たべました /tabemaʃita/ [*ate*] and はなして /hanaʃite/ [*speak IMP*]. After the experimental procedure, the first author conversed with the participant in Japanese, and she was then observed to delete vowels in these contexts in casual conversation suggesting that the participant was hyperarticulating during the experiment.

Finally, a difference was observed between the suffix environments, such that して /ʃite/ underwent deletion most frequently (*M* = 88%), followed by した /ʃita/ (*M* = 81%), and finally ます /masu/ (*M* = 78%). The small data set does not warrant statistical analysis, but we speculate that frequency of occurrence in Japanese is a likely explanation for this pattern. We thus counted the occurrences of each of the three suffixes in a balanced corpus of contemporary written Japanese [9], to examine whether there were differences in the frequency of the three suffixes in Japanese, consistent with the pattern observed in our data. In 105 million words, して /ʃite/ [*IMP*] occurred most frequently (1,104,381), followed by した /ʃita/ [*PST*] (860,509), and lastly ます /masu/ [*PRS*] (664,874). This suggests a relationship between vowel deletion and the frequency at which each suffix occurs, although a larger study is required to test this hypothesis more systematically.

Table 1. *Deletion and devoicing results. Words are presented in Hepburn Romanisation and environment indicates the environment of the vowel being monitored.*

/i/ Words	Environment	Deletion	Devoicing
shiteru	ʃ_t	89%	--
tabemashita	ʃ_t	92%	--
hanashite	ʃ_t	92%	--
ashita	ʃ_t	83%	--
shita	ʃ_t	67%	--
shite	ʃ_t	83%	--
shida	ʃ_d	--	--
shide	ʃ_d	--	--
mashi	ʃ_#	17%	--
nichi	C_#	25%	--
/u/ Words	**Environment**	**Deletion**	**Devoicing**
shuten	ʃ_t	33%	--
shute	ʃ_t	19%	--
suki	s_k	63%	--
daisuki	s_k	78%	3%
imasu	mas_#	67%	--
tabemasu	mas_#	83%	--
hanashimasu	mas_#	83%	--
karasu	s_#	17%	--
masu	mas_#	56%	--
moku	C_#	--	41%
/o/ Words	**Environment**	**Deletion**	**Devoicing**
kokoro	C_CV	3%	6%
hokori	C_CV	--	8%
/a/ Word	**Environment**	**Deletion**	**Devoicing**
haka	C_CV	--	--
Foils	**Environment**	**Deletion**	**Devoicing**
ima	--	--	--
ichiban	--	--	--
kasa	--	--	--
kaze	--	--	--
mizu	--	--	--
yasha	--	--	--
tusbasa	--	--	--

5. Discussion

The results of the present study show that Japanese vowel deletion is a frequently occurring phenomenon, even in words in citation form. Indeed, contrary to research that has found that devoicing does not occur in careful or slow speech, we show that Japanese speakers produce devoiced vowels also in (reasonably carefully monitored) citation form. While devoicing was rare for low vowels (e.g., こころ/kokoro/ [*heart*] 5.56%), ほこり/hokori/ [pride] (8.33%), and はか /haka/ [*grave*] (0%)), the only item with a high vowel in a devoicing environment, もく/moku/ [*wood*], elicited devoicing 41.67% of the time. These findings thus contrast with cross-linguistic data indicating that it is uncommon that speakers elide vowels in careful, planned speech, and the results are particularly interesting because of Japanese phonology's strict adherence to a syllable structure that does not allow for non-nasal consonant clusters.

Our results also show that deletion patterns in Japanese suffixes may be influenced by a number of factors. Of the suffix environments, して /ʃite/ [*IMP*] and した /ʃita/ [*PST*] elicited deletion more frequently than did ます/masu/ [*PRS*]. This might be the result of the /i/ being more susceptible to deletion than the /u/, the palato-alveolar fricative /ʃ/ being more likely to elicit deletion than the alveolar /s/, or the difference between medial and final vowels. However, all three suffixes were more likely to be deleted when attached to existing lexical items, indicating that there may simply be an implicit understanding that some lexical items require deletion or that the morphological status of the morpheme influences deletion. These possibilities are supported by the relatively rare deletion of /u/ in からす/karasu/ [*crow*] (17%), when compared to the ます/masu/ [*PRS*] elicitations (72%). This suggests the possibility that lexical frequency is a factor in the likelihood of deletion: vowels are more likely to be deleted in the three suffix environments because these occur very frequently in Japanese discourse [9].

Additionally, vowel deletion in the minimal pair して /ʃite/ [*IMP*] and しゅて /ʃute/ [*Nonce*] allows for the opportunity to examine the co-articulatory features of each vowel on the preceding fricative. Figure 2 shows the production of these two words by one of our participants. Interestingly, the F2 frequencies appear to be the same, these were measured at around 2200Hz which is to be expected with the high front vowel, /i/, but not the back vowel, /u/. There is, however, some F2 resonance lowering towards the end of the fricative in the しゅて /ʃute/ [*nonce*] example, which is likely the result of protrusion from lip-rounding. It should also be noted that the /ʃ/ in しゅて /ʃute/ [nonce] also achieved a higher intensity that the fricative in して /ʃite/ [*IMP*].

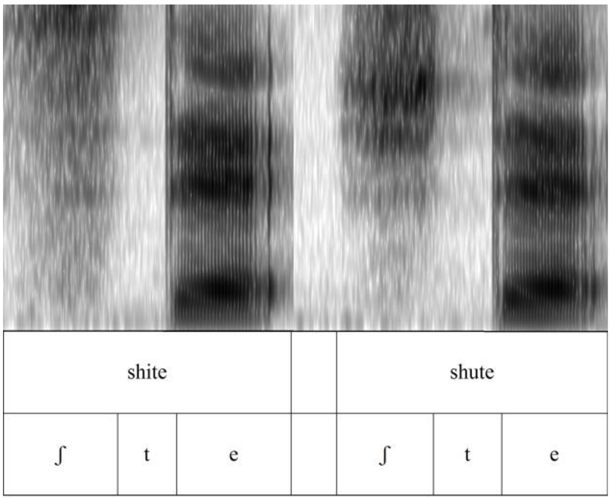


Figure 2: *Minimal Pair, して/ʃite/ [IMP] and しゅて /ʃute/ [Nonce].*

Finally, our data provides no support for previously reported regional differences in deletion/devoicing patterns in Japanese. Indeed, we found no significant difference between the behaviour of those from regions of Japan where reduction is common and those who are not. It is likely that this is the result of speakers from the Kansai and Shikoku regions adopting features of the higher prestige Standard Japanese due to the knowledge that their behaviour was being observed. Standard Japanese, the dialect spoken in centres of commerce and government, most closely resembles dialects spoken in and around Tokyo, a characterising feature of which is vowel reduction.

6. Conclusion

Deletion is generally regarded as a feature of fast, casual or connected speech. However, in Japanese, vowel deletion occurs when speakers produce planned, careful utterances. Vowel deletion and vowel devoicing are likely separate, although related, phenomena. Both behaviours occur in similar environments, deletion occurs after fricatives while devoicing occurs after plosives. While further research is required, another point of distinction is likely the relative frequency at which they occur. These behaviours can be distinguished by examining spectrographic evidence. While some may still argue that the vowel is masked by the noise of the preceding fricative, we present figure 1 which clearly shows that a devoiced vowel will result in an increase of frequency following a fricative.

We acknowledge that this is a preliminary report, but argue that frequency at which words occur in Japanese discourse appears to have a direct effect on the deletion of vowels. Those words containing the frequently occurring verb suffixes were more likely to elicit deleted vowels. This can be framed within the principle of least effort, which proposes that languages will move towards requiring less energy from the speaker. However, this tendency has to be balanced against the need for listeners to be able to differentiate between signifiers. The deletion of vowels in each suffix does not lead to homophony because the vowels have low informativeness—they are predictable and have a near-zero entropy. Therefore, deleting these vowels does not result in the loss of information in the signal.

7. Acknowledgements

We would like to thank our Japanese participants as well as those who helped to organise participants: Eri Suzuki and Takako Ikeda. We would also like to thank two anonymous reviewers for helpful suggestions and comments.

8. References

[1] Fujimoto, M., “Vowel Devoicing”, in H. Kubozono [Ed], Handbook of Japanese Phonetics and Phonology, Walter de Gruyter, 2015.

[2] Labrune, L., The Phonology of Japanese. Oxford University Press, 2012.

[3] Vance, T. J., An Introduction to Japanese Phonology, New York Press, 1987.

[4] Shibatani, M., The Languages of Japan, Cambridge University Press, 1990.

[5] Sugito, M., Nihongo Onsei no Kenkyu 3: Nihongo no oto, Izumi Shoin, 1996.

[6] Tsujimura, N., Introduction to Japanese Linguistics, Wiley Blackwell, 2013.

[7] Itô, J. and Mester, R. A., “The Phonology of Voicing in Japanese: Theoretical Consequences for Morphological Accessibility”, Linguistic Inquiry, 17(1), 49-73, 1986.

[8] Tsuchida, A., “Fricative-vowel Coarticulariion in Japanese devoiced syllables: Acoustic and Perceptual Evidence”, Working Papers of the Cornell Phonetic Laboratory, 9, 183-222, 1994.

[9] National Institute for Japanese Language and Linguistics, Kotonoha: Balanced Corpus of Contemporary Written Japanese. Retrieved from www.kotonoha.gr.jp/shonagon, 2009.

The Role of Nuclear Stress in Intelligibility: The Case of Cantonese Speakers of English

Simon Ka Ngai Leung[1], *Janice Wing Sze Wong*[2]

[1]English Language Teaching Unit, The Chinese University of Hong Kong, Hong Kong
[2]School of Communication, Hong Kong Baptist University, Hong Kong
simonleungkn@cuhk.edu.hk, janicewong@hkbu.edu.hk

Abstract

The current study investigates the effects of nuclear stress on intelligibility for listeners who are Cantonese speakers of English. Three groups of participants listened to different versions of the same speech: group A listened to the lecture with accentuation on new information; group B listened to the lecture with accentuation on given information; group C listened to the lecture with no accentuation. Results found that group A recognized and identified significantly more new information than group B and group C, suggesting the importance of nuclear stress production and placement to the communication between Cantonese speakers of English.

Index Terms: nuclear stress, intelligibility, Cantonese speakers of English

1. Introduction

English is now a language for communication both among speakers from various countries and between speakers in one country; this spread in the use of English has resulted in English as an International Language (EIL) [1]. Yet, there exists language variation among speakers owing to various reasons; when the variety is restricted to pronunciation, it is known as accent, which can be a part of their identities. To strike a balance between being intelligible and allowing them to convey their identities, Jenkins [2] proposed the Lingua Franca Core (LFC), which promotes both intelligibility and regional appropriateness among EIL interlocutors. The only suprasegmental item of the main core features in LFC is the production and placement of nuclear stress, which refers to the greatest prominence on a word or syllable in a clause or utterance [3]. Nevertheless, little research has empirically examined the extent to which nuclear stress patterns affect intelligibility.

The present study aims to investigate to what extent different nuclear stress patterns affects intelligibility of interlocutors who are Cantonese speakers of English. The three nuclear stress patterns under investigation were accentuation on new rather than given information, accentuation on given rather than new information, and absence of accentuation, which are the main nuclear stress patterns of varieties of English found in World Englishes research [4, 5]. Studies on 'intelligibility' [6] commonly define it as the listener's ability to recognize words or utterances, and assess it by having the participants transcribe the actual words or utterances in standard orthography while listening to the excerpt. However, since it is practically unrealistic for the listeners to transcribe all words in a discourse played at natural speed, added to the fact that listeners tend to pay most of their attention on new information in comprehending the meaning of utterances in a discourse, the current study precisely defined 'intelligibility' as the listener's ability to recognize and identify new information. This construct was assessed by having participants identify and transcribe new information.

2. Design and methodology

2.1. Research design

The present study adopted a between-group design, in which three groups of participants listened to different versions of the same speech which are identical except for differences in nuclear stress placement: group A listened to the lecture with accentuation on new rather than given information; group B listened to the lecture with accentuation on given but not new information; group C listened to the lecture with deaccentuation on both new and given information.

Participants in each group listened to one of the versions and responded by identifying and transcribing new information.

The main independent variable under investigation was the nuclear stress placement. The dependent variable was the number of tokens of new information identified and transcribed.

2.2. Stimuli

2.2.1. Materials

The current study adapted three versions of the speech from Hahn [7,8] which in turn was adapted from a naturally occurring academic lecture on individualism and collectivism. The difference of these three versions is illustrated as follows:

A: You want longer breaks but I want shorter breaks.
B: You want longer breaks but I want shorter breaks.
C: You want longer breaks but I want shorter breaks.

Hahn [7,8] categorized all ideas of the speech into main ideas and supporting details by asking 10 experienced English teachers to analyze and eliminate half of the message units which were not essential to the overall meaning of the speech; the message units eliminated by at least half of the raters were considered supporting details whereas the retained message units were classified as main ideas.

In preparation of Version A of the speech, Hahn [7,8] recorded the reading aloud of the text by seven English speakers, after which two pronunciation experts listened to their recordings to verify the nuclear stress placement. As for Version B of the text, three pronunciation instructors listened to the recordings of 20 English speakers from Outer Circle or Expanding Circle countries to verify the nuclear stress placement. Based on the nuclear stress placement of these speakers, an idealized and prototypical placement of nuclear stress was obtained and assigned to each message unit.

2.2.2. Talker

The first author was the talker. His primary languages at home, at school, and with friends are Cantonese and English. He is a suitable talker as Cantonese is a language which does not usually use nuclear stress; hence he has an internalized idea of how speech is produced without nuclear stress.

2.2.3. Stimuli preparation

The talker recorded the three versions of the speech in a research laboratory, using a Sanako SLH07 Professional Headset. The talker's utterances were digitally recorded using the audio recording software Praat [9] to save the sound files in .wav before they were archived.

While recording, he read aloud the stimuli at natural speed and attempted to maintain the speech rate throughout. After recording, the audio recording software Audacity was utilized to slice portions of stimuli together to make up the experimental stimuli, to ensure that the three versions of the stimuli were identical except for the nuclear stress placement, and to adjust the speech at an average rate [10].

After the stimuli were prepared, two other researchers were invited to listen to the recordings individually, and to verify that the intended nuclear stress placement in the three versions could be perceived. They were proficient Cantonese speakers of English, and possessed knowledge of English phonetics and phonology. There was a 98.6% agreement between them.

2.3. Participants

60 Cantonese speakers of English (12 male and 48 female; aged from 18 to 21) were randomly assigned to one of the three experimental groups. A post-experiment questionnaire revealed that they have Cantonese and English as their primary languages at home, at school, and with friends. All of them obtained at least a level 5 in the English Language subject of the internationally recognized Hong Kong Diploma of Secondary Education examination; therefore they were comparably proficient speakers. They also reported to have no hearing or speaking deficit. The three groups had no significant difference in gender [$\chi2(2) = 2.500$, $p = .287$] and in prior familiarity with the topic of the speech [$\chi2(2) = .436$, $p = .804$]. They participated in this experiment voluntarily.

2.4. Procedures

Participants were tested in a research laboratory at a local university. The participants were first introduced with the instructions and the format of the experiment, followed by answering some listening comprehension questions of the speech, rating its comprehensibility, as well as commenting on the speech and their reaction to it. Prior to testing intelligibility, in order to reach a consensus on how new information was defined, participants were provided with a short written text on which they had to identify and underline all the new information, after which the first author checked the answers with them. Then, they listened to the speech during which they identified its new information and wrote them down in standard orthography. Subsequent to the experiment, participants filled out a post-experiment language background questionnaire.

2.5. Data collection and analysis

To devise the key for data analysis, two teaching assistants at the Department of English, who are proficient Cantonese speakers of English, were invited to identify the new information of the text. Before that, to reach a consensus on how new information was defined, they were provided with the identical written text shown to participants before listening to the speech on which they identified and underlined all new information, after which the first author checked the answers with them. Next, they read the text of the speech, and identified and underlined its new information. If the words were chosen by these two teaching assistants and the first author, they were considered the key for data analysis.

In marking the scripts, for each transcribed word which is identical to the key, one point was awarded. All transcriptions were marked together twice by the first author. The time interval between the two markings was one month. The number of tokens of recognized and identified new information from main ideas, supporting details, and all ideas were then collected.

3. Results

3.1. Recognizing and identifying new information from all ideas

Table 1 shows a medium effect size in the scores between group A and group B, as well as between group A and group C, but a small effect size in the scores between group B and group C.

Table 1. *Descriptive statistics and effect sizes: Number of tokens of recognized and identified new information from all ideas for each group.*

Group	Number	Mean[#]	Standard Deviation	Effect size
A	20	20.95	3.252	.441 (A-B)
B	20	17.40	3.939	.012 (B-C)
C	20	17.30	4.508	.421 (A-C)

#. Possible maximum scores = 41

Figure 1 displays the boxplot of the interquartile range for each experimental group in recognizing and identifying new information from all ideas. The middle 50% of the scores shows the most stable results of the groups. It can be observed that the middle 50% of the scores of group A falls considerably higher than that of either group B or group C. Further, the middle 50% of the scores of group B is slightly higher than that of group C.

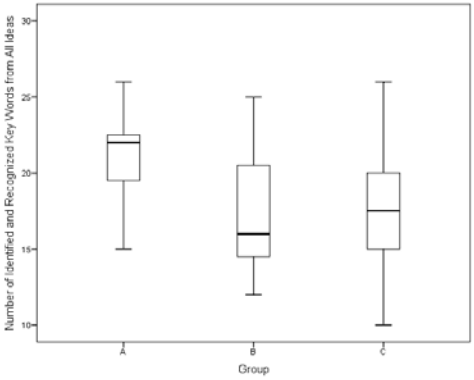


Figure 1: *Median and interquartile range of number of tokens of recognized and identified new information from all ideas for each group.*

Table of Contents for this manuscript

A one-factor analysis of variance (ANOVA) was then conducted to determine whether any of the differences was statistically significant. A significance level 0.05, two-tailed, was set for this and subsequent inferential statistics. A significant difference in the number of new information from all ideas among the three groups [*F*(2, 59) = 5.587, *p* = .006] was observed. Post-hoc pair-wise comparisons using Tukey's Honestly Significant Difference (HSD) among the groups revealed that, regarding the number of new information from all ideas, group A was significantly higher than that of group B (*p* = .016) as well as that of group C (*p* = .013) but no statistical difference was observed between group B and group C (*p* = .996).

This result reveals that group A, which listened to the speech with accentuation on new information but not on given information, recognized and identified more new information from the speech than either group B or group C. This result implies that nuclear stress did affect a listener's ability to recognize and identify new information from all ideas; in particular, nuclear stress placement on new rather than given information led to a listener's higher ability to recognize and identify new information from all ideas.

3.2. Recognizing and identifying new information from main ideas

Table 2 shows a medium effect size in the scores between group A and group B, as well as between group A and group C, but a small effect size in the scores between group B and group C.

Table 2. *Descriptive statistics and effect sizes: Number of new information from main ideas recognized and identified for each group.*

Version	Number	Mean[#]	SD	Effect size
A	20	17.55	2.605	.346 (A-B)
B	20	15.40	3.202	.076 (B-C)
C	20	14.85	3.990	.372 (A-C)

#. Possible maximum scores = 26

Figure 2 shows the boxplot of the interquartile range for each experimental group in recognizing and identifying new information from main ideas. It can be observed that the middle 50% of the scores of group A falls considerably higher than that of group C. In addition, the middle 50% of the scores of group A is slightly higher than that of group B which in turn is slightly higher than that group C.

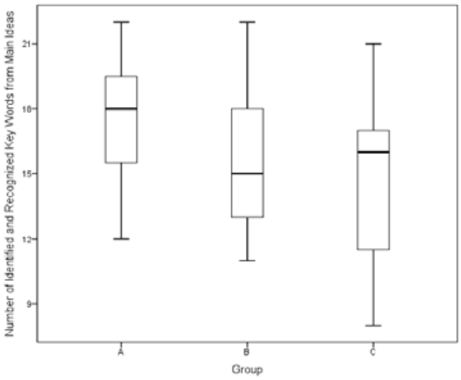


Figure 2: *Median and interquartile range of number of tokens of recognized and identified new information from main ideas for each group.*

An ANOVA was conducted to determine whether any of the differences were statistically significant. A significant difference in the number of new information from main ideas among the three groups was observed [*F*(2, 59) = 3.706, *p* = .031]. Post-hoc pair-wise comparisons using Tukey's Honestly Significant Difference (HSD) among the groups revealed that, regarding the number of new information from main ideas, group A was significantly higher than that of group C (*p* = .033) but no statistical difference was observed between group A and group B (*p* = .109), as well as between group B and group C (*p* = .860).

This result shows that group A, which listened to the speech with accentuation on new information rather than on given information, recognized and identified more new information from main ideas than group C. However, it seems that no significant difference was observed between group A and group B, as well as between group B and group C, in recognizing and identifying new information from main ideas. This result indicates that nuclear stress did affect a listener's ability to recognize and identify new information from main ideas. In particular, when comparing with no accentuation, nuclear stress placement on new rather than given information led to a listener's higher ability to recognize and identify new information from main ideas.

3.3. Recognizing and identifying new information from supporting details

Table 3 shows small effect sizes among all three groups. It appears to show that no remarkable difference exists among the three groups regarding the number of tokens of recognized and identified new information from supporting details. The negative effect size between group B and group C denotes that the mean score of group B is lower than that of group C.

Table 3. *Descriptive statistics and effect sizes: Number of tokens of recognized and identified new information from supporting details for each group.*

Group	Number	Mean[#]	SD	Effect size
A	20	3.50	2.544	.345 (A-B)
B	20	2.00	1.338	-.111 (B-C)
C	20	2.35	1.755	.254 (A-C)

#. Possible maximum scores = 15

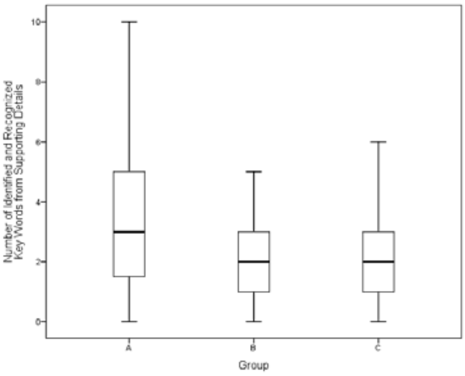


Figure 3: *Median and interquartile range of number of tokens of identified and recognized new information from supporting details for each group.*

Figure 3 above shows the boxplot of the interquartile range for each experimental group in recognizing and identifying new information from supporting details. It can be observed that in general the middle 50% of the scores of all groups are at comparable positions, representing the comparable number of tokens of new information from supporting details identified and recognized for all three groups.

An ANOVA was conducted to determine whether any of the differences were statistically significant. The results show a statistically significant difference in the number of tokens of new information from supporting details among the three groups [$F(2, 59) = 3.257$; $p = .046$], meaning that the differences found among the means on this variable were likely owing to the nuclear stress placement. Post-hoc pair-wise comparisons using Tukey's HSD revealed that, while no significant difference in the number of tokens of new information from supporting details identified and recognized was observed between group B and group C, as well as between group A and group C, a significant difference exists between group A and group B (p =.046). This result suggests that nuclear stress on given information might have lowered the listener's ability to recognize and identify words in supporting details.

4. Discussion

The present study shows that the participants could recognize and identify more new information in the speech with accentuation on new information but deaccentuation on given information than those with other stress patterns. This result implies that nuclear stress does affect a listener's ability to recognize and identify new information; in particular, nuclear stress placement on new information rather than given information leads to a listener's higher ability to recognize and identify new information. The results can be attributed to three factors: familiarity with accents, information processing, and listener attitudes.

The first factor is about the participants' familiarity with accents. Since the participants in the current study were more frequently exposed to accents of Inner Circle varieties due to their educational background and frequent exposure to English media of Inner Circle varieties, which is also in line with previous research findings [11], they were more familiar with the Inner Circle varieties of English which use accentuation on new information but deaccentuation on given information. Hence, participants could be considered to be more familiar with nuclear stress patterns of Version A. Such familiarity might explain the higher ability to recognize and identify new information for participants in group A [12].

Information processing of participants might also explain the results. Perhaps due to their familiarity with varieties of English from Inner Circle countries, it is likely that they expected nuclear stress to be applied on new information. The listeners might have used this knowledge to interpret the unaccented words as given information when hearing words with deaccentuation. Hence, in group B and C, as the speaker introduced new information with deaccentuation, the listener might have interpreted them as given information, which might have hindered their ability to recognize and identify new information. This explanation was evident among participants in group B and group C who commented that the stress placements failed, were wrong, or should be improved. It appears that they might have spent extra cognitive effort to notice the unexpected nuclear stress placements; to reinterpret the information status by using lexical, syntactic, and semantic clues; and to re-/identify the new information [7].

The present result could also be attributed to listener attitudes. From the open-ended responses, it was observed that the participants in groups B and C displayed more negative attitudes towards the speech, which are in the nuclear stress placement of varieties of English that deviate from Inner Circle varieties. Such negative attitudes, which in turn could have been owing to their familiarity with accents and information processing, also evident in previous attitudinal studies [13] might have undermined their ability to recognize words [14].

5. Conclusion

The current study reveals that nuclear stress production and placement is important for intelligibility in the communication between Cantonese speakers of English. The results have shed light on the extent to which phonological features affect intelligibility for the present target group and appear to have verified the importance of nuclear stress production and placement on intelligibility of Cantonese speakers of English, supporting part of Jenkins' Lingua Franca Core [2]. Future research, so as to fully verify the importance of nuclear stress placement in the LFC, should use speakers and listeners with different language backgrounds, English language proficiency, and roles, as well as speech with different genres, topics, and lengths.

6. References

[1] McKay, S. L., "Teaching English as an International Language: Rethinking goals and approaches", Oxford University Press, 2002.

[2] Jenkins, J., "The phonology of English as an International Language", Oxford University Press, 2000.

[3] Halliday, M. A. K., "Intonation and grammar in British English", Mouton & Co, 1967.

[4] Bolton, K., and Kwok, H., "The dynamics of the Hong Kong accent: Social identity and sociolinguistic description", J. Asian Pac. Comm., 1(1):147-172, 1990.

[5] Juffs, A., "Tone, syllable structure and interlanguage phonology: Chinese learners' stress errors", Inter. Rev. App. Ling., 28:99-117, 1990.

[6] Smith, L. E., and Rafiqzad, K., "English for cross-cultural communication: The question of intelligibility", TESOL Quarterly, 13(3):371-380, 1979.

[7] Hahn, L. D., "Native speakers' reactions to non-native stress in English discourse", unpublished PhD dissertation, 1999.

[8] Hahn, L. D., "Primary stress and intelligibility: Research to motivate the teaching of suprasegmentals", TESOL Quarterly, 38(2):201-223, 2004.

[9] Boersma, P., and Weenink, D. (2007). Praat (Version 4.5.25) [Online]. Available: http://www.praat.org, 2007.

[10] Rivers, W. M., "Teaching foreign language skills", University of Chicago Press, 1981.

[11] Evans, S., "Hong Kong English and the professional world", World Englishes, 30(3):293-316, 2011.

[12] Bent, T., and Bradlow, A. R., "The interlanguage speech intelligibility benefit", J Acou. Soc. Am., 114(3):1600-1610, 2003.

[13] Rubin, D. L., "Nonlanguage factors affecting undergraduates' judgments of nonnative English-speaking teaching assistants", Res. Higher Edu, 33(4):511-531, 1992.

[14] Brown, G., "Prosodic structure and the given/new distinction", in A. Cutler and D. R. Ladd [Eds], Prosody: Models and measurements, Springer-Verlag, 67-77, 1983.

Durational and Spectral Differences in Thai Diphthongs and Final Glides

Phongphat Man-khongdi[1], *Chutamanee Onsuwan*[1,2], *Charturong Tantibundhit*[2,3]

[1]Department of English and Linguistics, Thammasat University, Thailand
[2] Center of Excellence in Intelligent Informatics, Speech and Language Technology and Service Innovation (CILS), Thammasat University, Thailand
[3]Department of Electrical and Computer Engineering, Thammasat University, Thailand

phtmankhongdi@gmail.com, consuwan@tu.ac.th, tchartur@engr.tu.ac.th

Abstract

Acoustic analysis was conducted to compare Thai monophthongs /i, ii, ɯ, ɯɯ, u, uu, a, aa/, diphthongs /ia, ɯa, ua/, and vowel-to-glides (vowel + /j/ or /w/) in terms of duration, formant frequency, and spectral rate of change (TL_{roc}). Preliminary results from multiple repetitions of 30 target monosyllabic words show not only that Thai diphthongs and vowel-to-glides differ in their articulatory trajectories, but they appear to differ in their duration and TL_{roc} values. Average duration of diphthongs is shorter than that of vowel-to-glides. TL_{roc} of diphthongs is on average higher than that of long-vowel-to-glides, but lower than short-vowel-to-glides.

Index Terms: acoustics, diphthongs, final glides, Thai

1. Introduction

Despite the fact that vowels and glides sound quite similarly, it is widely accepted that phonetically, vowels and glides are distinct from one another. Glides do not maintain a steady state, but vowels (including diphthongs) do [1] and [2]. Some acoustic differences have been noted between diphthongs and glides. Acoustically, diphthongs and glides are investigated by analyzing a transition portion of two target sounds. The transition of diphthongs is usually slower and more gradual than that of vowel-to-glides [3]. This could be due to the fact that the production of glides moves in and out very quickly. Kenyon and Jones (1924) [4] stated that "rapid movement establishes /j, w/ as a set of 'gliding consonants' resulting from an immediate rapid movement of the lips and the tongue, or tongue alone" [5]. Gimson (1962) [6] also supported this idea and added that "a semivowel is a rapid vocalic glide on to a syllabic sound of greater steady duration" [4].

Of interest here are the diphthongs and final glides of Thai. Phonologically, Thai has three diphthongs, /ia/, /ɯa/ and /ua/. Unlike the 9 monophthongs which are phonemically contrastive in length, the diphthongs are not. However, their durations are predictable when followed by different finals [7], [8], [9], and [10]. Phonetically, /ia/, /ɯa/ and /ua/ could all be classified as opening diphthongs, that is, the onset is one of high vowels (/i/, /ɯ/, /u/), and the offset is low vowel (/a/). It should be noted that /i/, /ɯ/, /u/ and /a/ are among Thai monophthongs with long vowel counterparts. As for Thai glides, there are /j/ and /w/, each occurs in word initials and finals. Compared with initial glides, the final glides exhibit more limited co-occurrences with the vowels [11].

Only a handful of phonetics studies have been devoted to Thai diphthongs and glides. Among them, Roengpitya [9] investigated Thai diphthongs, /ia/, /ɯa/ and /ua/ in terms of duration and found that duration of the onset was longer than the offset in all diphthongs; duration of the offset was a crucial acoustic cue leading to the short-long difference. Later, Roengpitya [10] also examined duration as well as F1 and F2 at the 25%, 50% and 75% time points of Thai diphthongs. She concluded that durational differences among the diphthongs were related to syllable types. Moreover, in syllables where diphthongs were followed by /w/, F2 values were likely to increase at the 75% point, and where diphthongs were followed by /j/, F2 values are likely to decrease at the 75% point. In those two studies, however, differences between the Thai diphthongs and glides were not explicitly compared.

Tingsabadh and Abramson noted that when following monophthongs or diphthongs, /j/ and /w/ sound similar to /i/ and /u/, respectively [8]. Especially, when each of the two glides follows the low front vowel /a/ or /aa/; such sequences are in opposite directions to Thai (opening) diphthongs. On this ground, it is interesting to investigate whether and to what extent such vowel-to-glides differ acoustically from the diphthongs, and which acoustic characteristics could possibly set them apart. Another interesting point to explore is concerning Thai monophthongs and diphthongs. As previously mentioned, the onset (/i/, /ɯ/, /u/) and offset (/a/) of the three diphthongs are inherently monophthongs, it is worth investigating how much /i/, /ɯ/, /u/ and /a/ as part of diphthong trajectories differ from the monophthongs.

Therefore, the goal of this study is twofold. Firstly, we examine an extent to which the onset (/i/, /ɯ/, /u/) and offset (/a/) as part of diphthongs differ from when they occur as monophthongs. Secondly, we explore acoustic correlates which could be accounted for differences among Thai diphthongs and vowel-to-glides. Three acoustic measures: duration, F1 and F2, and spectral rate of change (TL_{roc}) are taken.

In the following sections, experimental method, main findings and discussions are given.

2. Method

2.1. Participants

Three male speakers of Thai (age range, 24-26 years) participated in this study. All of them were born and raised in the central region of Thailand and use Bangkok Thai in their daily life. They reported no known speech or hearing disorders and presented no speech production problems.

2.2. Word list

There were 30 target words in total. All are meaningful monosyllabic Thai words, begin with voiceless stop and have level tone (mid, low, or high). The words could be classified into 9 groups according to syllable types as shown in Table 1.

Each word was embedded in a carrier sentence [kam laŋ phûːt *target word* taːm pà ka ti] (I am speaking *target word* naturally) and read from a printout three times by each speaker. On the printout, the words were presented randomly in blocks. The last two repetitions of each were selected and analyzed.

Table 1: *30 target words (9 syllabic types) used in the study.*

Syllable types	Target words
Vʔ (short vowel with final stop)	tìʔ, pàʔ, ʔɯ̀ʔ, pùʔ, thɯ̀k
VV (long vowel with no final or final stop)	pii, paa, khɯɯ, tuu, pìik, pàak, pùut
Vw (short vowel with /w/)	tiw, taw
Vj (short vowel with /j/)	puj, kàj
VVw (long vowel with /w/)	taaw
VVj (long vowel with /j/)	taaj, kaaj
VD (diphthong with no final or with final stop)	pia, píaʔ, phɯ̀a, tua, ʔúaʔ, pìak, thɯ̀ak, pùat
VDw (diphthong with final /w/)	piaw
VDj (diphthong with final /j/)	pɯ̀aj, pùaj

2.3. Recording

All of the data were recorded by a recorder, MP3 Samsung model YP-Q2ABin a soundproof room at the Faculty of Liberal Arts, Thammasat University, Tha Phrachan Campus. Each recording session took about 10 minutes. The selected tokens were segmented and analyzed using PRAAT version 5.4.15 [12].

2.4. Measurements

Three measures were taken: duration of vowel (and final consonant), formant frequency (F1 and F2), and spectral rate of change (TL$_{roc}$) [13]. To obtain reliable comparisons among the vocalic portions (in ms) of monophthongs, diphthongs, and vowel-to-glides, final consonants (/w/, /j/, /t/, /k/ and /ʔ/) were included in the duration measure. It is worth noting that phonemically in Thai syllables with short vowel always ends with a final and that final stops are unreleased.

From the vocalic portion's duration, F1 and F2 values were manually extracted, at the 0%, 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, 90% and 100% points (11 points in total). Then, spectral rate of change (TL$_{roc}$) was calculated based on [13]. Specifically, the length of vowel section (VSL) was calculated for each of five separate sections, i.e., 30%-40%, 40%-50%, 50%-60%, 60%-70%, and 70%-80% using the formula

$$\mathrm{VSL_n} = \sqrt{(\mathrm{F1_n} - \mathrm{F1_{n+1}})^2 + (\mathrm{F2_n} - \mathrm{F2_{n+1}})^2}. \qquad (1)$$

Then, overall formant TL defined as a summation of trajectories of five vowel sections was calculated by

$$\mathrm{TL} = \textstyle\sum_{i=1}^{5} \mathrm{VSL_n}. \qquad (2)$$

The TL$_{roc}$ over the 50% portion of the vowel was calculated using

$$\mathrm{TL_{roc}} = \frac{\mathrm{TL}}{0.5\mathrm{V_{dur}}}. \qquad (3)$$

Finally, vowel section roc (VSL$_{roc}$) was calculated for each individual vowel section determined by the temporal location of the six measurement points, i.e., 30%-40%, 40%-50%, 50%-60%, 60%-70%, and 70%-80% using the formula

$$\mathrm{VSL_{roc_n}} = \frac{\mathrm{VSL_n}}{0.1\mathrm{V_{dur}}}. \qquad (4)$$

3. Results

The results are presented in terms of duration (section 3.1); values of F1 and F2 (section 3.2); and spectral rate of change (TL$_{roc}$) (section 3.3), as follows.

3.1. Duration

Figure 1 shows vocalic durations of the nine syllable types ranking from shortest to longest. Durational trend appears that VVw/j is the longest, followed by VDw/j, Vw/j, VV, VD, and Vʔ.

Durational ratio of long and short monophthongs is 2.2. Interestingly, average duration of diphthongs (VD = 227 ms) falls between, but is quite different from that of short and long monophthongs (Vʔ = 111 and VV = 248 ms).

For each diphthong, the average durations of /ia/, /ɯa/ and /ua/ are relatively comparable at 213.26, 228.49, and 239.32 ms respectively.

Short and long vowel-to-glides (Vw/j and VVw/j) are clearly longer than diphthongs (VD). Interestingly, when vowels are followed by glides, durational differences between short and long vowels become smaller and the whole portions (vowel + glide) are noticeably longer.

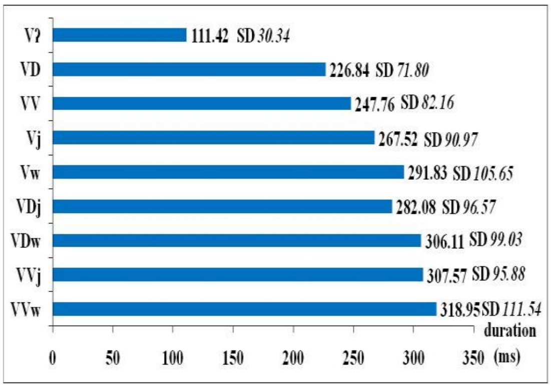


Figure 1: *Average duration of monophthongs and diphthongs, without any finals, or with final stops or final glides (Abbreviations as noted in Table 1).*

3.2. Formants

Figure 2 shows average F1 and F2 values for short and long monophthongs (at 50% point), diphthongs (solid lines showing downward trajectory at 0%, 50%, and 100% points), and vowel-to-glides /aj/, /aaj/, /aw/, and /aaw/ (dotted lines showing upward trajectory at 0%, 50%, and 100% points) and /iw/, /uj/ (dashed lines showing backing and fronting trajectories at 0%, 50%, and 100% points). Different types of vowel-to-glide are presented here to show an extent to which they are affected when preceded by low (dotted lines) versus high vowels (dashed lines).

It is clear that short versus long vowels of the same monophthong are relatively similar in terms of F1 and F2 (an exception might be for the /u/-/uu/ pair).For the three opening diphthongs, we could observe spectral changes in F1 and F2 values, specifically more so for F2 in the cases of /ua/ and /ia/. The diphthong onsets (/i/, /ɯ/, /u/ at 0%) and offsets (/a/ at 100%) as part of diphthongs appear to be more centralized than when they occur as monophthongs (exception might be for /ɯ/).

Similarly, onsets (/a/, /i/, /u/ at 0%) and offsets (/j/, /w/ at 100%) of the vowel-to-glides are relatively centralized.

3.3. TLroc

Spectral changes (movement) could be observed by F1-F2 plot (Figure 2), but TL$_{\text{roc}}$ values certainly give us a clearer comparison when the length of vowel section is individually taken into account (See Section 2.4).

Figure 3 shows TL$_{\text{roc}}$ values of the nine syllable types ranking from lowest to highest. The values seems to suggest the trend with VDw/j being the highest followed by Vw/j, VD, VVw/j, Vʔ, and VV.

As expected, short and long monophthongs have the lowest TL$_{\text{roc}}$ values. TL$_{\text{roc}}$ of diphthongs (VD = 5.5 Hz/ms) lies between, but is quite different from that of short vowels with final glide and long vowels with final glide (Vw/j = 6.2 and VVw/j = 3.6 Hz/ms).

Separately, TL$_{\text{roc}}$ values of each diphthong vary from /ɯa/ (2.49 Hz/ms), /ua/ (6.36 Hz/ms) to /ia/ (6.66 Hz/ms).

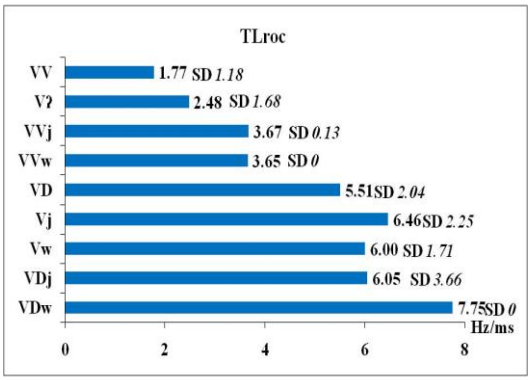


Figure 3: *TLroc values of monophthongs and diphthongs, without any finals, or with final stops or final glides (Abbreviations as noted in Table 1).*

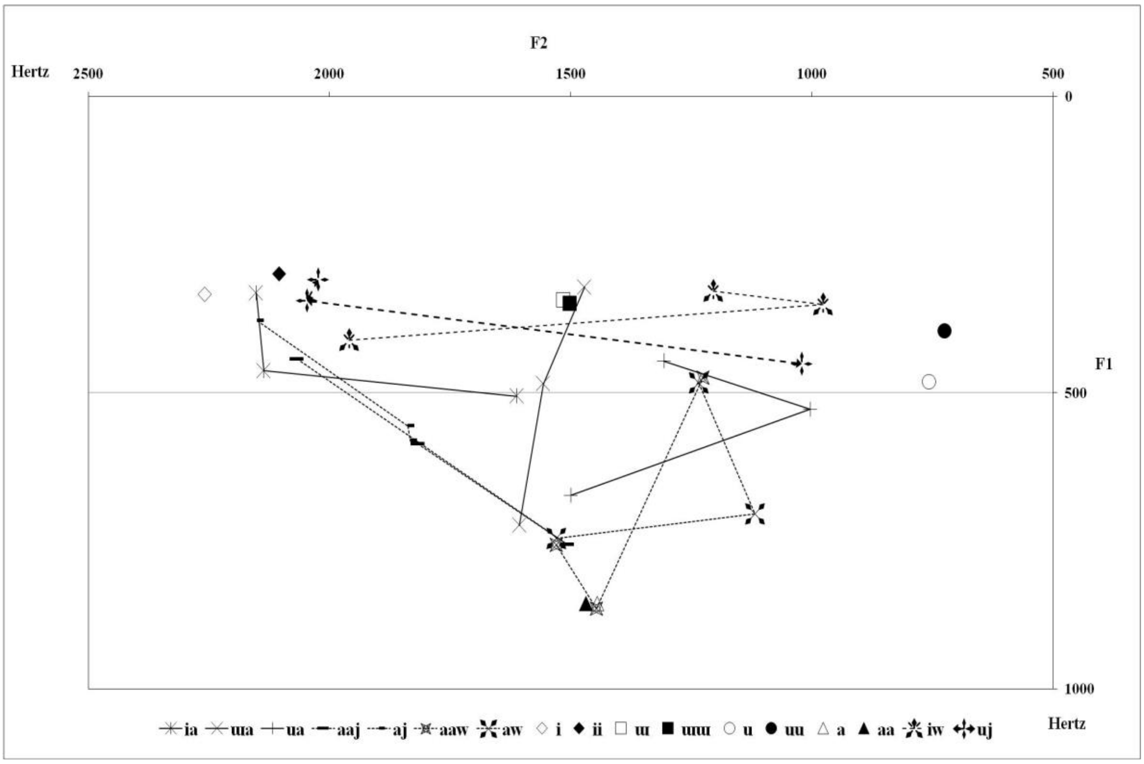


Figure 2: *Average F1 and F2 values of short and long monophthongs (at 50% point), diphthongs (solid lines showing downward trajectory at 0%, 50%, and 100% points), and vowel-to-glides /aj/, /aaj/, /aw/, /aaw/ (dotted lines showing upward trajectory at 0%, 50%, and 100% points) and /iw/, /uj/ (dashed lines showing backing and fronting trajectories at 0%, 50%, and 100% points).*

4. Discussions and future work

Durational values of short and long monophthongs as well as F1 and F2 range reported here are in agreement with previous studies of Thai [7]. The findings show that for the most part Thai diphthongs and vowel-to-glides are relatively different in terms of duration, F1 and F2 values, and spectral rate of change. Principally, in the case of Thai, differences in duration and spectral rate of change seem to be more relevant. Diphthongs are relatively shorter than either short or long monophthongs with glide. As noted by Roengpitya [9] and [10], Thai syllables with final glide appear to be longer than with any other types of final consonant (nasals and stops) and durational differences between short and long vowels become much less. This seems to be the case here, although direct comparison among different syllable types with various types of final consonant could provide a clearer picture.

In terms of F1 and F2 values for the Thai diphthongs, F2 dimension seems to capture spectral changes more clearly than F1, particularly for /ua/ and /ia/. Moreover, the diphthong onsets (/i/, /ɯ/, /u/) and offsets (/a/) as part of diphthongs appear to be more centralized than when they occur in monophthongs.

Lastly, TL$_{roc}$ values, which have been used in a few studies of diphthongs and glides of other languages [13], [14] and [15], seem to be useful and reliable for capturing spectral dynamics of diphthongs and vowel-to-glides in this current study. TL$_{roc}$ of Thai diphthongs is found on average to be higher than that of long-vowel-to-glides, but lower than short-vowel-to-glides. It remains to be seen with larger acoustic data if our finding agrees with the statement saying that the transition of diphthongs is usually slower and more gradual than that of the vowel-to-glide [3].

It is noteworthy that degrees of intensity were included in some studies of diphthongs and glides [16] and [17]. However, our preliminary acoustic analysis failed to show any noticeable differences in intensity level among Thai diphthongs and glides.

Finally, our future direction is to further explore acoustic properties of Thai diphthongs and glides from more speakers and to conduct detailed statistical analysis.

5. References

[1] Kent, R. D. and Read, C. *The acoustic analysis of speech* (2 ed.). Singular Publishing Group, 2002.

[2] Raphael, L. J., Borden, G. J. and Harris, K. S. *Speech science primer: physiology, acoustics, and perception of speech* (5 ed.). Lippincott Williams and Wilkins, 2006.

[3] Olive, J. P., Greenwood, A. and Coleman, J. *Acoustics of American English Speech: a dynamic approach*. New York: Springer-Verlag, 1993.

[4] Kenyon, J. S. In: Lance, D.M., Kingsbury, S.D. (Eds.), *American Pronunciation*. George Wahr Publishing Company, Ann Arbor, MI, 1924. (Citation from expanded 12th edition, 1997).

[5] Maddieson, I. "Glides and germination," *Lingua*, 118, 1926-1936, 2008.

[6] Gimson, A. C. *An Introduction to the Pronunciation of English*. London: Edward Arnold, 1962.

[7] Abramson, A. S. *"The vowels and tones of standard Thai: Acoustical measurements and experiments*," Indiana U. Research Center in Anthropology, Folklore, and Linguistics, Pub. 20. Bloomington, 1962.

[8] Tingsabadh, K. and Abramson, A. S. "Thai," *Journal of the International Phonetic Association*, 23:1, 1993.

[9] Roengpitya, R. "A Study of Vowels, Diphthongs, and Tones in Thai." Ph.D. dissertation. University of California at Berkeley, 2001.

[10] ___________. "Different Durations of Diphthongs in Thai: a New Finding," *The Proceedings of the Annual Meeting of Berkeley Linguistics Society*, 2002.

[11] Munthuli, A., Sirimujalin, P., Tantibundhit, C., Kosawat, K. and Onsuwan, C. "A Corpus-based Study of Phoneme distribution in Thai," *Proceedings of the 10th International Symposium Natural Language Processing (SNLP), Phuket, Thailand*. 114-121, 2013.

[12] Boersma, P. and Weenink, D. "Praat: doing phonetics by computer [Computer program]. Version 5.4.15," Amsterdam: University of Amsterdam, 2015.

[13] Fox, R. A. and Jacewicz, E. "Cross-dialectal variation in formant dynamics of American English vowels," *The Journal of the Acoustical Society of America,* 126 (5), 2603-2618, 2009.

[14] Asu, E. L., Lippus, P., Niit, E. and Türk, H. "The Acoustic Characteristics of Monophthongs and Diphthongs in the Kihnu Variety of Estonian," *Linguistica Uralica*, XLVIII, 3, 2012.

[15] Mayr, R. and Davies, H. "A cross-dialect acoustic study of the monophthongs and diphthongs of Welsh," *Journal of the International Phonetic Association*, 41, 1-25, 2011.

[16] Mauder, E. and Van Heuven, V. "On the rise and fall of the Spanish diphthongs," In C. Cremers, & M. Den Dikken (Eds.), *Linguistics in the Netherlands*, p.171-182. Amsterdam: John Benjamins, 1996.

[17] Jaggers, Z. S. "Acoustic cues to the [j]-[i] distinction in American English," *Talk presented at the 170th Meeting of the Acoustical Society of America (ASA 170)*. Jacksonville, 2-6 November, 2015.

Effect of Clinical Depression on Automatic Speaker Verification

Sheeraz Memon[1], *Mukhtiar Ali Unar*[1] *and Bhawani Shankar Chowdhry*[2]

[1]Department of Computer System Engineering, Mehran UET, Jamshoro, Pakistan
[2]Department of Electronic Engineering, Mehran UET, Jamshoro, Pakistan
{sheeraz.memon, mukhtiar.unar, bhawani.chowdhry}@faculty.muet.edu.pk

Abstract

The effect of a clinical environment on the accuracy of the speaker verification was tested. The speaker verification tests were performed within homogeneous environments containing clinically depressed speakers only, and non-depressed speakers only, as well as within mixed environments containing different mixtures of both clinically depressed and non-depressed speakers. The speaker verification framework included the MFCCs features and the GMM modeling and classification method. The speaker verification experiments within homogeneous environments showed 5.1% increase of the EER within the clinically depressed environment when compared to the non-depressed environment. It indicated that the clinical depression increases the intra-speaker variability and makes the speaker verification task more challenging. Experiments with mixed environments indicated that the increase of the percentage of the depressed individuals within a mixed environment increases the speaker verification equal error rates.

Index Terms: *Speaker verification, Clinical environment, Clinical depression.*

1. Introduction

The performance of speaker recognition systems degrades due to both, the intra-speaker variability and the background noise. One of the factors affecting the intra-speaker variability is the clinical depression. Speech contents of clinically depressed speakers consist of more abstractive flow of conversations, higher frequency of pauses and more nonverbal sounds than speech of normal speakers [19]. It has been also previously demonstrated that clinical depression changes acoustic characteristics of speech [1-3, 13-15] and therefore, it can be hypothesized that the speaker verification accuracy can be affected in an environment consisting fully or partially of clinically depressed people.

This paper aims to determine the effects of clinical environments consisting of clinically depressed people on the speaker verification rates when using the state of the art speaker recognition techniques. The importance of this study is given by the fact that robust speaker recognition systems have potential applications in the clinical environments and the health care sectors such as telemedicine, biometrics and surveillance systems. It is recently reported that nearly 20 percent of military service members who have returned from Iraq and Afghanistan report symptoms of post-traumatic stress disorder or major depression, according to a new RAND Corporation study [22].

The state of the art speaker recognition systems extracts acoustic features which capture the characteristics of the speech production system such as pitch or energy contours [7], glottal waveforms [6], or formant amplitude and frequency modulation [5] and model them using statistical learning techniques [20,21]. The *Mel frequency cepstral coefficients* (MFCCs) have been commonly used to characterize acoustic properties of speech [8] often in conjunction with the *Gaussian mixtures model* (GMM), which is regarded to be one of the best statistical modeling techniques used in speaker recognition systems [8,16,17,18]. The characteristic features used in this study include MFCCs, their velocity and acceleration, short time energy and zero crossing rates. The pre-processing stage was used to separate the silence/noise intervals and to perform the pre-emphasis filtering. The speaker models were built using the Gaussian mixture modeling based on the expectation maximization (EM) procedure [16,17].

The remaining part of this paper is organized as follows. Section 2 describes the speaker verification system. In Section 3, the experimental setup and results are presented, and finally, Section 3 contains the conclusions.

2. Speaker Verification System

2.1 General Framework

The general framework of the speaker verification system used to conduct our experiments is shown in Figure.1. The system can operate in one of the three possible modes: universal background model (UBM) training mode, target speaker enrollment mode and testing mode. In each case identical speech detection and feature extraction methods are used. An energy based silence detector was used to discard the low energy intervals of the signal [16]. Previous research has shown that the MFCC based systems are not very sensitive to changes in frame size (in the range 20-50ms) and frame step (in the range 1/6 to 1/3 of the frame size). Frames whose energy is too low to be considered speaker-discriminative were therefore excluded from subsequent processing. From each remaining frame, the first 12 MFCC were computed and normalized using the cepstral mean subtraction (CMS) method.

The sequences of feature vectors were then modeled with the GMM. For each target speaker 1024 Gaussian mixtures were generated. Each model was defined by a set of parameters including its *a priori* probability, mean vector, and the diagonal covariance matrix. The speaker models were trained with around 5 minutes of data length for each speaker. After the enrollment (training) stage, the universal background model (UBM) parameters [18] were derived using the expectation maximization (EM) algorithm trained on a large speech corpus including the non-target speakers obtained from the NIST 2001 and NIST 2002 SRE corpora. The target speaker's model means were then adapted using the maximum a posteriori (MAP) estimation method, the UBM and the target speaker's data. During the testing stage, the same pre-processing and feature extraction methods were applied to the test data as in the training stage. The testing sequences of feature vectors were then scored by each speaker's model, and the verification decision was made based on the identity of the highest scoring model. The general system performance was assessed using the equal error rate (EER) measure and by plotting the detection error trade-off (DET) curves.

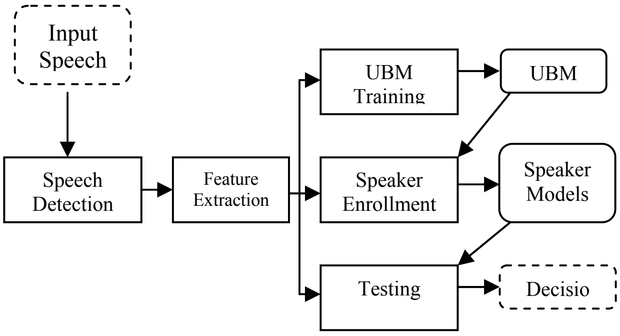


Figure.1. *General framework of the speaker verification system.*

2.2. Clinical Speech Corpus

The clinical data used in this study (CC-ORI) was obtained from the Oregon Research Institute, USA. The non-clinical data was obtained from the NIST 2004 corpus.

The CC-ORI consists of audio recordings of 139 adolescents (93 females and 46 male) aged 13-19, participating in typical discussions between family members. Each subject was represented by around one hour of recording. Details of the data acquisition sessions can be found in [9-11]. Through the self-report and interview measures of the adolescence's depression evaluated by psychologists from ORI [9], 68 (49 females and 19 males) were diagnosed as suffering from major depressive disorder (MDD), and the remainder (44 females and 27 males) were healthy controls (i.e., no current or lifetime history of MDD). The clinically depressed and healthy groups were matched on their demographic data which included their sex, race and age. The CC-ORI recordings were sampled with the 8 kHz rate and 16 bits/sample.

The NIST 2004 samples were derived from the Switchboard 2, and the Mixer projects. The Switchboard-2 Corpora included mostly college or early post-college age students [4] from a specific area of the United States. The NIST 2004 data was also sampled at 8 kHz rate with 16bits/s.

2.3. Speech Segmentation and Feature Extraction

The speech signal was segmented using the Hamming window into short frames of length 20 ms within which the spectral and temporal properties of speech such as signal energy and pitch can be assumed stationary [8,12]. There was 50% overlap between frames. The feature extraction was performed on the frame-by-frame basis. Each frame was used to derive a feature vector consisting of: 12 MFCCs coefficients, 12 Δ-MFCC (first derivative of MFCCs), 12ΔΔ-MFCC (second derivative of MFCCs), 1-short time energy coefficient and 1-zero-crossing coefficients. The resulting arrays of 38-dimensional feature vectors were used to test the speaker verification rates in different environments

3. Experiments and Results

3.1. Individual Class Speaker Verification (ICSV) within Homogeneous Environments using CC-ORI

In this experiment, the speaker verification was performed within two homogeneous environments. The first environment contained 100% of depressed speakers and the second environment contained 100% non-depressed speakers. The depressed environment contained 68 speakers (49 females and 19 males), and the non-depressed environment contained 71 speakers (44 females and 27 males). The experimental results for the intra-class speaker verification tests (ICSV) test within homogeneous environments are presented in Figure.2. It can be clearly observed that the speaker verification task within the depressed environment is more challenging than within the non-depressed environment. The speaker verification equal error rate (EER) for the depressed speakers is 5.1% higher than for the non-depressed speakers. Since the numbers of speakers in both clinically depressed and non-depressed classes were almost the same, and the utterances were recorded under the same background noise conditions, it can be concluded that the clinical depression was the main factor causing the degradation of speaker verification accuracy within the depressed environment when compared with the non-depressed environment. It also indicates that the intra-speaker variability within the depressed environment is higher than within the non-depressed environment.

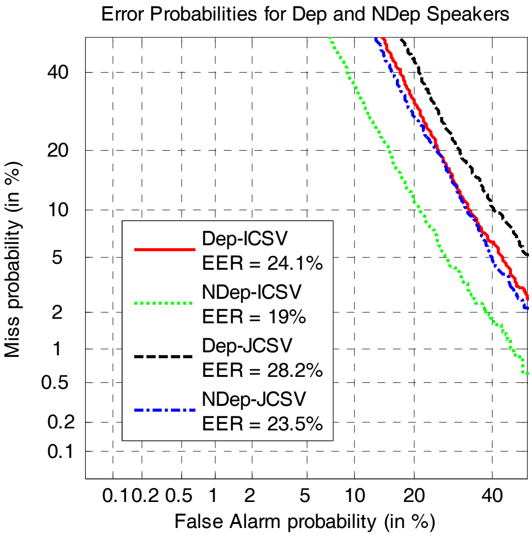


Figure. 2. *Detection Error Trade-Off (DET) curves and Equal Error Rates (EERs) for the ICSV test within homogeneous environments.*

3.2. Speaker Verification within Mixed Environments using CC-ORI

This set of experiment tested the speaker verification accuracy within mixed environments i.e., environments consisting of different mixtures of both, depressed and non-depressed speakers. The detection error trade-off (DET) curves and the equal error rates (EERs) for the mixed environments constructed out of the CC_ORI data are presented in Figure.3, denoted as Joint-class speaker verification (JCSV). Four different environmental mixtures were used. Each mixture contained a fixed number of 68 non-depressed speakers. The first mixture had no depressed speakers, the second mixture contained 17 depressed speakers, the third mixture contained 34 depressed speakers and the fourth mixture contained 68 depressed speakers. Since, the mixed environments were composed of speakers from the CI-ORI corpus only; the conditions of recordings and the background noise were the same for all speakers. Also, each environment contained the same number of non-depressed speakers; only the number of depressed speakers was changing. It can be therefore assumed that the observed effects on speaker verification were mostly due to different amounts of depressed individuals within a given environment. The results in Figure.3 show that the increasing percentage of the depressed speakers within a mixed environment leads to an increase of the EER values.

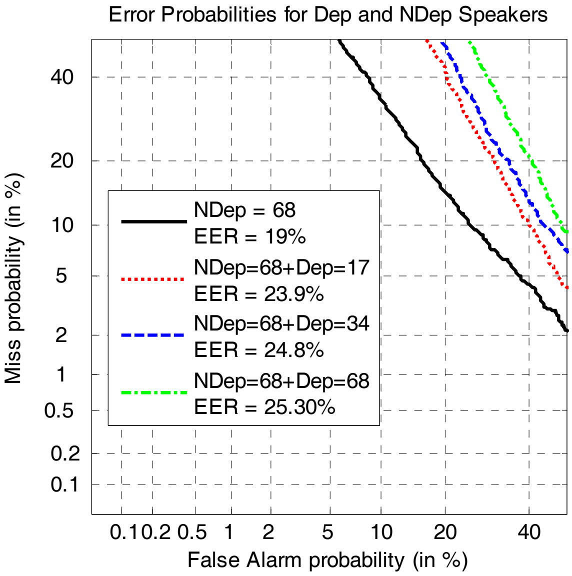


Figure.3. *Detection Error Trade-Off (DET) curves and Equal Error Rates (EERs) for the JCSV test within mixed environments using CC-ORI.*

3.3. Speaker Verification within Mixed Environments using CC-ORI and NIST2004

This set of experiment again tested the speaker verification accuracy within mixed environments however this time the depressed speakers were sourced from the CC-ORI and the non-depressed speakers were taken from the NIST 2004 corpora. The resulting detection error trade-off (DET) curves and the equal error rates (EERs) for the mixed environments constructed out of the CC_ORI data are presented in Figure.4. Three different environmental mixtures were used. The first mixture contained 68 depressed speakers from CC-ORI and no non-depressed speakers. The second mixture contained 616 non-depressed speakers from NIST 2004 and the third mixture contained 616 non-depressed speakers from NIST 2004 and 68 depressed speakers from CC-ORI.

Since, the mixed environments were composed of speakers from two different data bases (CI-ORI and NIST 2004), the recording conditions and the background noise were different. It is therefore difficult to draw any definite conclusions. Figure.4 shows that the addition of 68 depressed CC-ORI speakers to the 616 NIST 2004 non-depressed speakers increases the EER values compare to the environment containing only the 616 NIST 2004 speakers. This could be the result of both, the clinical depression and the different noise level in the CC-ORI recordings. To be able to draw more definite conclusions, further research on equalization methods compensating for the differences in the recording conditions is needed.

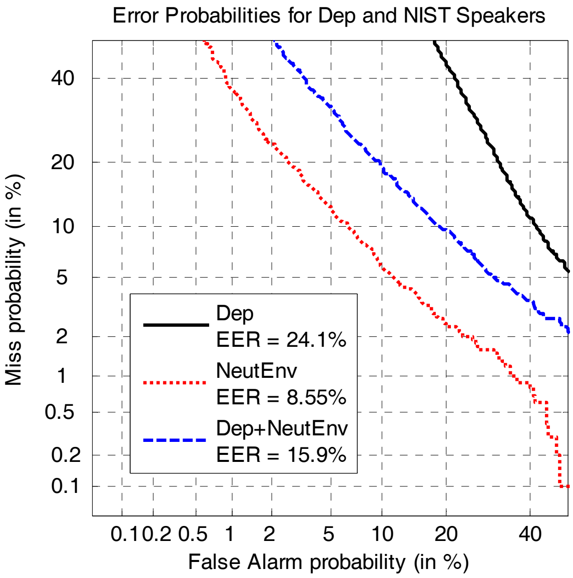


Figure.4. *Detection Error Trade-Off (DET) curves and Equal Error Rates (EERs) for the JCSV test within mixed environments using CC-ORI+NIST200*

4. Conclusions

The effects of the clinical depression on the speaker verification accuracy were investigated. The tests conducted within homogeneous environments using the same data base clearly indicated that the speaker verification within the clinically depressed environment is more challenging than within the non-depressed environment and the EER values obtained within the depressed environment are higher than within the non-depressed environment. The tests conducted within mixed environments composed of the same data base indicated that the higher is the percentage of the depressed speakers, the larger are the speaker verification EER values. Finally, the tests conducted within mixed environments constructed out of two different data bases were not conclusive due to the lack of equalization methods allowing to directly merge data recorded under different conditions.

5. Acknowledgements

The authors would like to thank the Oregon Research Institute, U.S.A. and Dr. Lisa Sheeber for providing the clinical data base. This study supported in part by the Australian Research Council Linkage Project LP0776235.

6. References

[1] Karlsson, I., Banziger, T., Dankovicova, J., Johnstone, T., Lindberg, J., Melin, H., Nolan, F., Scherer, K. “Speaker verification with elicited speaking styles in the VeriVox project,” Speech Communication 31(2-3): 121-129, 2000.

[2] K. R. Scherer, T. Johnstone, G. Klasmeyer, & T. Bänziger “Can automatic speaker verification be improved by training the algorithms on emotional speech” University of Geneva, Switzerland.

[3] Murray, I. R., & Arnott, J. L. "Toward a simulation of emotion in synthetic speech: A review of the literature on human vocal emotion," JASA, 93: 1097-1108, 1993.

[4] Alvin F. Martin, "Encyclopedia of Biometrics" National Institute of Standards and Technology Gaithersburg, Maryland, USA.

[5] C. R. Jankowski jr. et al., "Fine structure features for speaker identification," in Proc. ICASSP, 1996, pp. 689–692.

[6] M. D. Plumpe, T. F. Quatieri, and D. A. Reynolds, "Modeling of the glottal flow derivative waveform with application to speaker identification," IEEE Trans. Speech Audio Process., vol. 7, no. 5, pp. 569–586, Sep. 1999.

[7] B. Peskin et al., "Using prosodic and conversational features for high performance speaker recognition: Report from JHU WS02," in Proc. ICASSP, vol. 4, 2003, pp. 792–795.

[8] D. A. Reynolds, "Experimental evaluation of features for robust speaker identification," IEEE Trans. Speech Audio Process., vol. 2, no. 4, pp. 639–643, Oct. 1994.

[9] L. Sheeber, H. Hops, J. Andrews, T. Alpert, and B. Davis, "Interactional processes in families with depressed and non-depressed adolescents: reinforcement of depressive behaviour," Behaviour Research and Therapy, vol. 36, pp. 417-427, 1998.

[10] H. Hops, A. Biglan, A. Tolman, L. Sherman, J. Arthur, and N. Longoria, "Living in family environments (LIFE) coding system: Reference manual for coders," Oregon Research Institute, Eugene, OR, Unpublished manuscript, 2003.

[11] H. Hops, B. Davis, and N. Longoria, "Methodological issues in direct observation-illustrations with the living in familial environments (LIFE) coding system," Journal of Clinical Child Psychology, vol. 24, pp. 193-203, 1995.

[12] L.R. Rabiner, R.W. Schafer, "Digital Processing of Speech Signals", Prentice-Hall Inc.

[13] D. J. France, et al. "Acoustical properties of speech as indicators of depression and suicidal risk," IEEE Transactions, Biomedical Engineering, vol. 47, pp. 829-837, 2000.

[14] E. Moore, et al. "Critical analysis of the impact of glottal features in the classification of clinical depression in speech," IEEE Trans, Biomed Eng, vol. 55, pp. 96-107, 2008.

[15] A. Ozdas, et al. "Investigation of vocal jitter and glottal flow spectrum as possible cues for depression and near-term suicidal risk," IEEE Trans, Biomed Eng, vol. 51, pp. 1530-1540, 2004.

[16] Reynolds, D. A., Rose, R. C., and Smith, M. J. T., PC-based TMS320C30 implementation of the Gaussian mixture model text-independent speaker recognition system. In Proceedings of the International Conference on Signal Processing Applications and Technology, November 1992, pp. 967–973.

[17] D. Reynolds and R. Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models," in *IEEE Trans. Speech and Audio Processing*, 1995, vol. 3, pp. 72–83.

[18] D. Reynolds, T. Quatieri, and R. Dunn, "Speaker verification using adapted Gaussian mixture models," in *Digital Signal Processing*, 2000, vol. 10, pp. 19–41.

[19] Christopher. D and Marybeth. S," Intra-speaker variability in palatometric measures of consonant articulation", Journal of Communication Disorders, Volume 42, Issue 6, Dec 2009, Pages 397-407.

[20] Liping Chen, Kong Aik Lee, Bin Ma, Wu Guo, Haizhou Li and Li Rong Dai, "Local variability vector for text-independent speaker verification", *9th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, pp. 54-58, 2014.

[21] Patrick Kenny; Gilles Boulianne; Pierre Ouellet; Pierre Dumouchel, "Speaker and Session Variability in GMM-Based Speaker Verification " , IEEE Transactions on Audio, Speech, and Language Processing,, Volume: 15, Issue: 4 Pages: 1448 - 1460, 2007.

[22] The masks of war: American military styles in strategy and analysis: A RAND Corporation research study, C Builder, Johns Hopkins University Press.

Using version control to facilitate a reproducible and collaborative workflow in acoustic phonetics

Nay San

Macquarie University

`nay.san@mq.edu.au`

Abstract

This paper outlines the benefits of using a version control system (VCS) in an acoustic phonetics workflow. Selected features of Git, a VCS, are introduced and explained in terms of their relevance to facilitating communication and reproducible research, especially in projects with multiple contributors. A workflow, as used in two current projects on Australian Aboriginal languages, is provided as an illustrative example. While the paper focusses on the Git VCS with Praat as the annotation tool, the workflow may be easily adopted using other VCS tools (e.g. Mercurial) and annotation tools (e.g. ELAN, EMU).

Index Terms: collaborative annotation, reproducible research, corpus management

1. Introduction

1.1. Role of annotations in speech corpora

A large part of any acoustic phonetic project lies in deriving linguistically relevant information from recorded audio or video signals. The information derived, for example, may be orthographic or phonetic transcriptions, temporal locations of segment boundaries, or pitch/formant tracks.

Ultimately, the analysis of such annotations generally constitute the results of interest, and various platforms have been proposed to facilitate reproducible analyses. For example, Alveo [1] provides the Galaxy workflow engine, through which analyses may be published as, essentially, an interactive flowchart. This allows for each analysis step to be reproduced on demand.

While producing documented and reproducible analysis workflows are highly important, the same aspects in annotation workflows have received little attention—despite the fact that analyses depend heavily on these annotations.

Annotation workflows can often be equally complex multi-step and multi-level processes. Consider a single aspect, automation. Creating annotations may involve some degree of automation (fully-, semi-automated, or fully-manual). For example, transcriptions may involve using only speech-to-text software (fully-automated), its use with supervision from a human transcriber (semi-manual), or rely solely on human transcription (fully-manual). Additionally, a given project may also choose to employ a mix of methods to create the desired annotations. For example, word-level transcriptions may be first manually transcribed, while segment-level boundaries are then derived via forced-alignment.

Given that annotation processes can be complex, it is very likely that a given set of annotations contains a number of errors. Indeed, annotation errors are well documented even in large, publicly-released corpora [2]. Thus, annotation data cannot be expected to remain entirely static, as errors should be corrected as they are discovered.

However, for analyses to be reproducible, the annotations used in the analyses must not change. Thus, it appears that annotation data are required to be both static and dynamic, at the same time. The use of version control, however, can satisfy this seemingly paradoxical requirement.

1.2. Collaborative workflows

Projects may also involve several annotators in differing locations. One way to handle this complexity has been to establish a central annotation server [3], where users create and modify annotations through a web application. This solution, however, may not be appropriate where contributors do not have frequent and/or fast access to the server (e.g. those in the field).

The common solution for collaboration appears to be to use a cloud hosting service, such as Dropbox or Google Drive, to synchronize annotation files (.TextGrid, .ELAN, etc.). While these services do offer some level of version control (revert to old versions), multiple contributors editing a single file often results in conflicting copies of the files (e.g. `fileA.TextGrid`, `fileA (1).TextGrid`). Not only do most cloud hosting applications lack explicit notification of conflicts as they emerge, they do not provide a convenient method for resolving them—especially when the conflicting files have to be merged together.

Thus, collaboration in annotation workflows not only require a way to share annotation files, but also to merge the resulting annotations from the various collaborators.

1.3. Overview

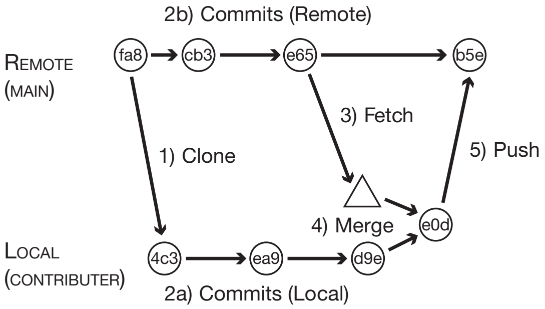
The annotation of speech data from audiovisual signals can be a complex, cyclical and distributed process. Facilitating this process thus requires a system for tracking, documenting and synchronising changes. Additionally, the system should also allow for changes to be made locally using standard annotation software (e.g. Praat [4], EMU [5]).

Through a description of Git, a version control system (VCS), this paper will show that a VCS is exactly what is required to satisfy the requirements stated above. Section 2 provides an introduction to Git, first defining some key terminology in Git, then describing the relevance of some features of Git in the acoustic phonetics workflow (Subsections 2.1-2.4). Section 3 provides an illustrative example of two acoustic phonetics projects using Git in the their workflows. Finally, the benefits of using a Git VCS in the workflow are summarised, and potential issues are highlighted in the discussion (Section 4).

2. Git

Git is a VCS that was developed in 2005 by Linus Torvalds, the creator of the Linux kernel. As the Linux kernel is free and open source, its development is driven by contributions from a large number of individuals. This requires a system for allowing the development to be distributed among many contributors, a way to review these contributions, and easily merge various contributions together into the main project.

A key feature of the Git VCS is that the system may take various forms, and it is not constrained to one 'standard' workflow. Various projects may adopt one that suits them best, and integrate Git accordingly. Figure 1 provides an illustration of some steps involved in adding a contribution to a relatively simple project model.



Note. Each circle represents a set of registered changes, a 'commit', which are given a unique identifier (text inside circle, abbreviated to 3 characters)

Figure 1: *A simplified illustration of the Git contribution process*

Clone. In Step 1), the contributor *clones* the project from an external source onto their local computer (analogous to downloading the files).

Commit. In Steps 2a) and 2b), various changes are registered or *committed* to the project. Commits can be thought of as the registration of significant changes. A project can be reverted back to a specific commit at any point in the future (i.e. undo various changes). Notice that Steps 2a) and b) can be done in parallel, and creation of commits require no communication between the local and remote sources.

Fetch. In Step 3), the contributor *fetches* changes from the remote source. This allows the contributor to view all the changes that have taken place since the last synchronisation.

Merge. In Step 4), the contributor *merges* their local changes with those fetched from the remote repository. Conflicting changes between the local and remote sources may prevent a successful merge. The contributor must resolve these conflicts during the merging process.

Push. In Step 5), following a successful merge, the contributor's changes can be integrated into the main project by *pushing* these changes up to the remote source (analogous uploading the files).

2.1. Commits: registered sets of changes

As mentioned, commits consist of registering various changes to files with the Git VCS. While 'Track Changes' in a Microsoft Word document, for example, might show *what* changes were done (and by whom and when), the context for the change, or *why*, is not explicitly required. Though, the why is very often a key piece of information. Moreover, Git allows the logical grouping of changes according to such information.

Consider an illustrated commit below, shown in Figure 2. The difference, or 'diff', between consecutive versions of two TextGrid files indicate *what* changes have taken place. For instance, the annotator John Doe has edited an annotation from 'balap' to 'palap' in File1.TextGrid (using Praat), and this edit has resulted in a change on Line 20 of the file.

```
Commit:  ea9
Author:  John Doe
Date:    9 June 2016, 2:05 pm
Message: Stop voicing not contrastive. Updated
         transcription procedure uses voiceless
         allophones.

Diff from 4c3 to ea9:

File1.TextGrid
19 19   ...
20    - text = "balap"
   20 + text = "palap"
21 21   ...

File5.TextGrid
33 33   ...
34    - text = "golot"
   34 + text = "kolot"
35 35   ...
```

Figure 2: *An illustration of a commit (ID: ea9) and the diff 'difference' in two TextGrid files between commits 4c3 and ea9*

A similar edit has also seen a change on Line 34 of File5.TextGrid. Changes to the two files have been registered under a single commit, ea9—with the reason for the changes provided in the required commit message. Commits can thus entail not only single changes, but sets of changes across multiple files. Conveniently, Git allows both single files (e.g. only File1.TextGrid) or entire sets of changes (e.g. both files) to be easily reverted to previous versions.

Commits thus provide the primary means of tracking and documenting significant changes to the various files in a given project. The flexibility in reverting allows for commits to be either entirely, or only partially, undone in the future.

2.2. Tags: noteworthy commits

A specific commit may be 'tagged' with some additional metadata. In software development, this is generally used to flag released versions of software, e.g. v 1.1.5. Analogous to software releases, tags can be used to note the exact versions of annotation data used various research output (e.g. ICPHS2015, SST2016).

The tagged version can be 'checked out' easily by anyone wishing to access an older state of the data. Essentially, a given analysis can continue to access the state of the data at a specific timepoint even without reverting current project files to that timepoint.

2.3. Distributed sources, controlled synchronisation

Cloud filesharing services such as Dropbox or Google Drive—unless paused—propagate changes across all synchronised folders immediately when the changes are performed. While

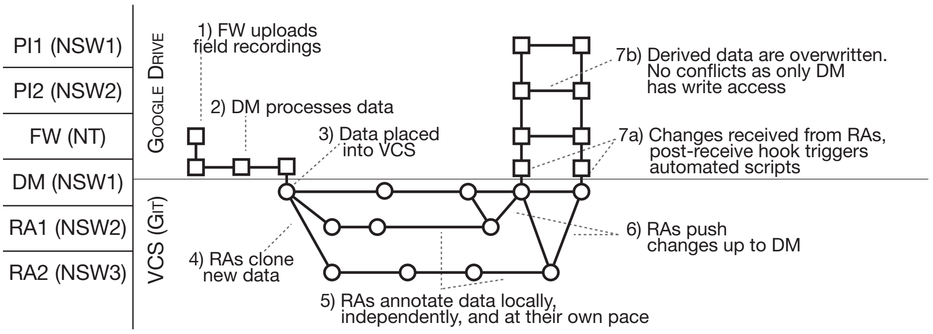

PI1 (NSW1)
PI2 (NSW2)
FW (NT)
DM (NSW1)
RA1 (NSW2)
RA2 (NSW3)
Google Drive
VCS (Git)
1) FW uploads field recordings
2) DM processes data
3) Data placed into VCS
4) RAs clone new data
5) RAs annotate data locally, independently, and at their own pace
6) RAs push changes up to DM
7a) Changes received from RAs, post-receive hook triggers automated scripts
7b) Derived data are overwritten. No conflicts as only DM has write access

Figure 3: *Annotated illustration of a workflow in the Alyawarre/Anmatyerre Dictionary Recordings (AADR) project. Data are shared by various parties: investigators (PI1, PI2), fieldwork sources (FW), data manager (DM), and research assistants (RA1, RA2) across multiple locations (NT, NSW1, 2, 3)*

in most cases, this may be desired, a level of control over which changes and when they are synchronised may be desired in others. Moreover, immediate synchronisation of changes may create several conflicting copies if two annotators need to work on the same file concurrently.

With the Fetch-Merge-Push procedure of Git, annotators may retrieve (fetch) and send (push) changes in a controlled manner. In the majority of cases, Git is able to automatically merge changes to one file from multiple sources. Should a merge fail, Git alerts the user of the exact file(s) and line(s) within the file(s) preventing the merge; making conflict resolution relatively straightforward.

Thus, the use of version control reduces the overall clutter in project directories, eliminating not only the need for keeping manually-versioned files (e.g. 20160608-data-final-1.csv), but also conflicting copies from arising in shared directories.

It should also be noted that the use of Git does not prevent the use of services such as Dropbox or Google Drive. A portion of the project files may lie in a version controlled folder, while others lie in a cloud sync folder, e.g. in Dropbox. The illustrative example below will demonstrate such a use case.

2.4. Hooks: triggers for automation

As certain Git actions are associated with notable changes in the state of various files, Git hooks provide a mechanism of triggering follow-up actions when certain Git actions occur. For instance, a post-commit hook is an executable script that is run by the Git VCS after a user has created a new commit. As Git is language-agnostic, these scripts may be written in any language (e.g. Python, R).

While such hooks may be used for a variety of purposes (e.g. e-mail alerts, data validation), they provide a means of monitoring the state of a set of annotation data. For example, a simple script generating proportions of .wav files with and without corresponding .TextGrid files can provide a rough estimate of completed annotations. As the script is triggered after every commit, these proportions update automatically.

Additionally, monitoring annotation data may involve some simple extraction or data transformation procedures, e.g. from nested to tabular data. The benefits of generating such data early in the annotation process are two-fold. Firstly, since each commit generally consists of small sets of changes, annotation errors produced in the immediate or recent set of commits can be detected and sourced easily, as well as fixed promptly. Secondly, the extraction and transformation scripts can be repurposed for the data wrangling scripts in various analyses.

In other words, leveraging automated scripts to monitor annotation data can significantly reduce the considerable amount of time that is often spent validating and transforming data into appropriate forms prior to most analyses.

3. Illustrative example

Kaytetye, Alyawarre and Anmatyerre are three languages from the Arandic family, spoken in the region surrounding Alice Springs, NT. A core part of the Kaytetye Phonology (KPHON) and Alyawarre/Anmatyerre Dictionary Recordings (AADR) projects involve establishing data repositories of (mainly) audio recordings around the three languages. The data from these repositories are to be used in generating language resources (e.g. multimedia dictionaries), and to establish quantitative accounts of various phonological structures and processes.

The dictionary recordings in AADR consist of various speakers reading aloud the dictionary headwords in citation form, with several repetitions per headword. One goal across the two projects is the phonetic transcription of all the dictionary words by at least two transcribers. The transcription process is blind, in that annotators are not given the identity of the words, orthographic or otherwise.

Figure 3 illustrates an example of an annotation workflow in the AADR project; from receiving fieldwork recordings to the generation of the first set of derived data. In Step 1), fieldwork recordings are uploaded to a shared Google Drive. In Step 2), the data manager (DM) splits the lengthy recording sessions (e.g. 20160503-Wa-We.wav) into individual words (e.g. wampe.wav) for archiving. In Step 3), the single word files are anonymised (e.g. wampe.wav to file07.wav). These files are then received by the research assistants (RAs) for transcription (Step 4). As these files are cloned by the RAs through Git, all changes made by the RAs are tracked through Git commits (Step 5). When a portion (or all) files have been annotated, RAs may send the changed files back to the DM (Step 6).

When changes are received by the DM, a post-receive hook runs a number of scripts (Step 7), producing various reports. Table 1 displays the first 4-lines of data from such a automatically generated report. There are three columns: vowel, n (tally), prop (proportion), which lists the unique vowels in the annotation data, the number of each vowel, and its relative proportion in the data set. The data indicates, at the time of generation, there have been 376 transcriptions of word-medial unstressed [ɘ], and this accounts for 14.75% of the data set. Unstressed word-final [ɘ] (Row 3) accounts for 10.63% of the data.

vowel	n	prop
ə	376	14.75
ˈɐ	336	13.18
ə#	271	10.63
ɐ	260	10.20
...	...	...

Table 1: *Initial 4 rows of a table summarising the vowel data in the repository, generated automatically after commits are received*

The regular, *automated* derivation of summary data, allows for changes in the annotations to be closely monitored. If Table 1 were reverse-ordered on the column n, the data would show the least-frequently occurring transcriptions. From this view of the data, transcription errors become immediately obvious. For example, suppose there is single instance of ‘y’ in this table (i.e. a high front rounded vowel). Given the languages being analysed, it is highly likely that it is an error (perhaps a typo of the adjacent key ‘u’).

Additionally, the writing of code snippets for summarising the dataset assists in the data analysis process. Any future analysis requiring the extraction of vowel annotations from this data set can re-purpose the extraction parts of the scripts being used to monitor the data.

Moreover, the data extraction script need not be written in any executable language. As Git hooks are simply events that trigger a script, the scripts themselves may be in any language—be it an emuR query, a Praat or Python script—and can thus help automate repetitive tasks across a number of applications.

4. Discussion

Many of the ideas introduced and discussed are already part of the standard phonetics workflow. For example, there is usually some manual form of version control for files and the [manual] running of scripts to extract and verify data. Git as a VCS, along with hosts such as GitHub, formalises such ideas into a set of standard protocols, which allow for efficiency and transparency in the workflow—factors which become increasingly important with large-scale collaborative work.

The start up costs for implementing and using Git as a VCS are now quite minimal. This has been aided by the establishment of many cross-platform Graphical User Interface (GUI) clients: for example, GitHub Desktop, GitKraken, and SourceTree (used by all AADR and KPHON RAs). That is, regardless of the contributor’s platform (Windows, OS X, Linux), everyone can use the same Git client, and become acquainted to a Git VCS workflow as a group.

Additionally, using Git as a VCS does not limit a project to a certain workflow. This fact has allowed for the training of RAs in an incremental manner. The simplest possible workflow through a Git client merely adds a number of administrative steps in the client to send and receive data. Then, additional features are introduced one by one as needed.

For KPHON and AADR, the teaching of Git/SourceTree to RAs is part of the introduction to the projects. For example, when RAs initially start any project, they must familiarise themselves with the project’s annotation criteria on a practice set. For KPHON and AADR, the Git/SourceTree training is simply a part of this familiarisation.

Moreover, the formalisation of how problematic annotations and errors are communicated (e.g. via GitHub issues), in fact, facilitates the workflow, as it establishes a central repository of knowledge. For example, an RA can quickly find out if a specific problem has been encountered before by another, or past, RA, and how that problem was resolved (with details of the exact files concerned).

One should note, however, that Git and particularly its diff functionalities (i.e. summarising the changes between two versions of files) work best on non-binary data, i.e. plain text file formats (which include Praat .TextGrid, ELAN .eaf, .json). This is due to the fact that software development has depended primarily on version control of source code, and not binary assets such as image or audio files. However, as Git as a VCS is being adopted more and more by teams outside the software development tradition, more and more diff functionalities are being implemented by third parties. In short, versions of any file can be tracked through the Git VCS, however, quickly viewing what changes occured currently only works best for plain text file formats.

Similarly, automatic conflict resolution in Git works best for line-by-line mismatches in these plain-text files. This may be problematic for annotation files such as .TextGrids as their data consist of inter-related multi-line information (e.g. time + text). Such problems, however, would only emerge when a specific portion of the same tier of a given file is edited by two or more annotators at the same time. We have not encountered this as an issue in KPHON or AADR.

5. Conclusion

This paper introduced the use of a version control system (Git, in particular) in an acoustic phonetics annotation workflow. Of course, Git and other version control systems were designed with software development in mind. Consequently, it may well be discovered that—for use within acoustic phonetics—only a subset of their features are highly beneficial, while some unnecessary, and others lacking—and must be developed for the field of acoustic phonetics. For any such discovery to take place, however, gradual adoption, experimentation, and subsequent discussion of version controlled data and its role in in acoustic phonetics will be necessary.

6. References

[1] S. Cassidy, D. Estival, T. Jones, D. Burnham, and J. Burghold, “The Alveo virtual laboratory: A web based repository API,” in *9th Language Resources and Evaluation Conference (LREC 2014)*, Reykjavik, Iceland, 2014.

[2] A. Rosenberg, “Rethinking the corpus: Moving towards dynamic linguistic resources.” in *INTERSPEECH*, 2012, pp. 1392–1395.

[3] J. Poignant, M. Budnik, H. Bredin, C. Barras, M. Stefas, P. Bruneau, G. Adda, L. Besacier, H. Ekenel, G. Francopoulo, J. Hernando, J. Mariani, R. Morros, G. Quénot, S. Rosset, and T. Tamisier, “The CAMOMILE collaborative annotation platform for multi-modal, multi-lingual and multi-media documents,” in *10th Language Resources and Evaluation Conference (LREC 2016)*, Portorož, Slovenia, 2016.

[4] P. Boersma, “Praat, a system for doing phonetics by computer,” *Glot international*, vol. 5, no. 9/10, pp. 341–345, 2002.

Preservation of Tone in Right-Dominant Tone Sandhi: A Fragment of Disyllabic Tone Sandhi in Máodiàn Wú Chinese

Ruiqing Shen[1] *& Phil Rose*[2]
[1]Hong Kong University of Science & Technology
[2]ANU Emeritus Faculty
shenruiqings@hotmail.com, philjohn.rose@gmail.com

Abstract

Impressionistic and acoustic data are presented for the nine citation tones, and a small part of the disyllabic tone sandhi, of a speaker of the previously undescribed Chinese dialect of Maodian 毛店 from the Wuzhou 婺州 subgroup of Wu 吳. The data are used to refine the typology of the apparent right-dominant tone sandhi characteristic of the southern Wu and Min area. It is shown that not all word-final tones are the same as citation tones; and that therefore preservation of word-final tones cannot be criterial for right-dominance.

Index Terms: Tone Sandhi, Right-dominance, Tonal acoustics, Wu dialects, Wuzhou, Maodian.

1. Introduction

Language likes to exploit the polarity of metrical strength. One striking example is the typological difference, independent of segmental phonotactics, between right- and left-dominant tone sandhi systems found in the highly complex morphotonemics of the so-called *sandhi-zone* of China’s eastern coastal provinces [1, 2]. In right-dominant varieties, it is the tones on the morphemes on the rightmost syllables of a word which determine the sandhi shape. The tone on the word-final syllable is said to be 'preserved', 'unchanged' or 'in agreement with' the citation tone, and tonal contrasts on the preceding syllables tend to be neutralised, although the neutralisation groupings are often bewilderingly complicated. Right-dominant varieties are said to be found in the southern Wu and Min dialects [3, p.287], but the exact distribution is not known. The variety described in this paper is located the southern Wu subgroup of Wuzhou, so it can be expected to be right-dominant.

But is right-dominance monolithic, or are there degrees to this typological parameter? We present data from a variety in the right-dominant area which appear to show the latter. We will focus in this paper on just one aspect of the relationship described as criterial for right-dominant sandhi: the extent to which the citation tone values are preserved on word-final syllables. For the purposes of this paper we define *preservation* thus. Preservation occurs *iif* word-final and citation tonal acoustics are intrinsically related, i.e. if the word-final tonal acoustics can be understood to be the same as the citation tone, once allowance is made for the effect of occurrence in word-final position and the expected perseverative assimilatory effect from the tone on the preceding syllable.

Since the disyllabic data are complicated we only have space to present two cases: one (relatively) simple, which is prototypically right-dominant; and one which is clearly not. We describe the citation tones first, then tones on disyllabic words with the simple case preceding the more complex.

2. Data

The data are from a high quality recording which was part of a wider survey of tones and tone sandhi in ca. 30 southern Wu varieties conducted by Professor W. Ballard in 1988 [4], and his generosity in making the data available is gratefully acknowledged here. The recording consists of three replicates each of ca. 40 monosyllabic and 190 disyllabic utterances elicited from a then 27 year old male speaker who was born and grew-up in Maodian until 17. A comparison of Wu dialect descriptions done in 1928 and 1992 [5,6] shows that they changed considerably in this ca. thirty year period, and more recent socio-phonetic findings [7] suggest that tonal change is accelerating, at least in metropolitan areas. In a sense, therefore, this description may also be partly considered a salvage operation. The digitised recordings – both citation tones and disyllabic tone sandhi – can be listened to on the second author’s web-page (http://philjohnrose.net), where their individual and mean acoustics are also plotted.

3. Citation tones

In Chinese tonology, a citation tone is the tone given to a morpheme when its Chinese character, which may represent either a free or bound morpheme, is read out. The speaker has nine citation tones which may be described auditorily as follows (segmentals are transcribed phonemically). The **upper-mid level tone**, a reflex of Middle Chinese (MC) tone Ia**,** has a level pitch contour in the upper third of the pitch range, e.g. fi *fly* 飛, kɑ *liver* 肝, nuŋ *east* 東. The **lower-mid level tone**, from MC Ib, has level pitch in the lower third of the pitch range, e.g. bi *skin* 皮, dzɯa *tea* 茶, nia *year* 年. The **mid rising tone** (< MC IIa) has prolonged pitch in the mid pitch range with a final rise, e.g. siəu *arm* 手, nia *point* 點, hua *fire* 火. The **lower-mid rising tone** (< MC IVa) has the same delayed pitch rise a little below that of the mid rise tone e.g. siɛ *snow* 雪, ɓɛi *north* 北. The **low rising** tone (<MC IIb, IVb) has pitch which rises from low in the speaker’s pitch range to mid with a prolonged initial component, e.g. bi *blanket* 被, zua *sit* 坐 , [ɔ] *to study* 學, dɑu *poison* 蝳. The **high falling** tone (<MC IIIa) has pitch which falls through the speaker’s modal pitch range, e.g. si *four* 四, t^{h}ɯa *to jump* 跳. The **depressed high falling tone** (<MC IIIb) has similar pitch to the high falling tone, but with a low onset which results in a convex pitch contour in the bottom two thirds of the pitch range. Examples are di *ground* 地, vɑ *rice* 飯, miɛ *face* 面. The **short stopped mid tone** (<MC IVa) has a short pitch in the lower-mid pitch range truncated by a glottal stop, e.g. kuaʔ *bone* 骨, tɕ^{h}yaʔ *to come out* 出. The **short stopped low rise**

tone (<MC IVb) has a short rising pitch in the lower third of the pitch range truncated by a glottal stop, e.g. zaʔ *ten* 十.

This rather large number of observed tones relates primarily to the historical development of morphemes with tonal cognates of Middle Chinese so-called *entering tones* IVa and IVb. Originally, in Proto Wu say, these two tones had short duration and ended in a glottal stop. In many modern Wu dialects their reflexes retain these features and are still considered separate tones; but in other varieties the tones have lost their glottal stop and undergone further development. In some Wuzhou varieties the short tones have lengthened and merged with other tones; in others they have lengthened but remained separate by virtue of different pitch shapes [8, p.23]. Interestingly, the Maodian speaker provided a further variation on this theme, in that he clearly showed a merger of etymological tone IVb with tone IIb (the low rising tone), whilst keeping a lengthened version of etymological tone IVa separate (as the lower-mid rise tone).

This situation was further complicated, however, by a phenomenon, again said to be typical for Wuzhou, whereby some morphemes with etymological IVa and IVb tones have alternative phonological shapes [8, p.23]. One shape is conservative, preserving the short pitch ending in a glottal stop; the other is the innovative lengthened tone. This alternation was also shown by the Maodian speaker. For most IVa and IVb cognates he had innovative long reflexes. For a few IVa and IVb cognates, however, he retained a conservative short stopped tonal shape. Although this phenomenon is traditionally termed 文白異讀 *different colloquial and literary character readings*, there was nothing in the linguistic structure of any of the morphemes involved that would serve as an obvious conditioning factor. Thus, for example, he read the characters for the morphemes *bone*, *come out* and *ten* with short stopped tones, but, in the same formal elicitation session, those for *snow*, *put out* and *month* were given long tones. Indeed, Ballard's notes show some free variation, in that *bone* was also said with a long tone. Although the conditioning of such short forms remains elusive, therefore, it is clear that one has to deal with nine different tonal shapes.

Citation tone acoustics were quantified with the same method used in a previous study of a right-dominant Wu variety [9]. A wideband spectrogram was generated in *Praat*, together with its wave-form and superimposed F0. The token's tonally relevant F0 was then identified, extracted and modeled in *R* by an 8th order polynomial. This enabled F0 values to be sampled from the polynomial F0 curve with a sufficiently high sampling frequency (at 10% points of the curve as well as 5% and 95%) to capture the details of its time-course.

The mean tonal acoustics of the nine Maodian citation tones (F0 as function of duration) are shown in figure 1. The tonal F0 shapes have been plotted in two panels, as their complex configuration would have made it difficult to identify their shapes otherwise. In order to demonstrate that some IVb morphemes have indeed merged with reflexes of IIb, the low rising tone is plotted separately with a green dotted line for its IVb and a black dotted line for its IIb constituents. Their extreme similarity indicates provenance from the same synchronic tone.

The F0 shapes of the individual tones are clear and generally correspond fairly well to their pitch descriptions. The two short stopped tones (brown) can be seen to have a duration of about half that of the unstopped tones, and also to have very similar onsets to their corresponding long tones (green). The lower-mid level tone (blue) appears to have a slightly depressed onset extending for the first 10 csec. or so. One clear area of disagreement between the F0 and tonal pitch is in the high falling and depressed high falling tones (red). Their offsets, between ca. 105 Hz and 110 Hz, lie considerably below the two low rising tones that sound to lie near the bottom of the speaker's range. Including these falling tone offsets as tonally relevant will have the effect of distorting the way the F0 represents tonal pitch, and they are best considered as idiosyncratic offset perturbations (some speakers end their falling tones with a glottal stop or creak; others, like this Maodian speaker, have a gradual offset to modal phonation). In the following sections, we describe tone sandhi in words ending with morphemes which carry the lower-mid level and mid rising citation tone.

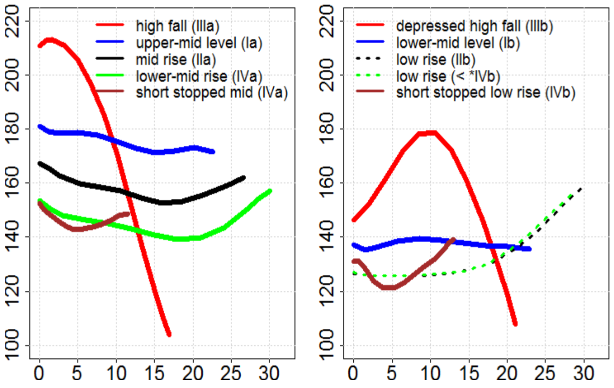


Figure 1: *Mean F0 (Hz) of the speaker's nine isolation tones plotted against mean duration (csec.).*

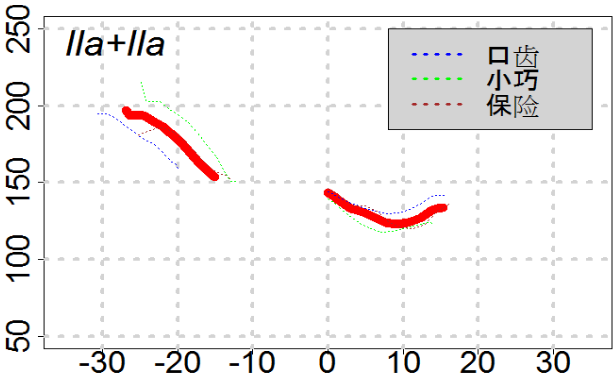


Figure 2. *Tonal acoustics of three Maodian words with mid rising morphotonemes on both syllables. Thick solid line = mean F0, dotted line = individual tokens' F0. X axis = duration (csec.) y axis = F0 (Hz).*

4. Disyllabic tone sandhi: procedure

The same procedure for extracting the tonal acoustics of the disyllabic words was used as in [9], which sampled F0 as a function of the word's segmental structure – first Rhyme, intervocalic consonant and second Rhyme – using 8th order polynomial modeling in *R*. Three different words were measured for each etymological tonal combination, and their mean values calculated. Figure 2 shows the individual values and their mean for three words with the mid rising morphotoneme on both syllables. These have a high falling pitch on the first syllable followed by a low dipping pitch on the word-final syllable. (One example was /ɓɯa ɕie/ [52.212] *insure*, which is a synonym compound consisting of the bound morpheme {保 *protect* ɓɯa 323} and the free morpheme {險 *danger* ɕie 323}. F0 is plotted as a function of absolute duration aligned at onset of second-syllable Rhyme (csec.0). The fairly tight clustering of the individual words' F0 values is typical.

5. Disyllabic tone sandhi: a minimally complex example

To demonstrate the mechanics of the least complex tone sandhi in Maodian disyllabic words, we examine combinations with underlying mid rising tone on the word-final syllable, and all tones on the preceding syllable. Examples are given in table 1. The procrustean Chao five-point scale transcribing tonal pitch is intended as convenient abbreviation only.

Table 1. *Examples of Maodian speaker's tone sandhi in words with underlying mid rising tone on word-final syllable. Pitch representations are color-coded with figure 3.*

word-final mid-dipping isolation tone [323] preceded by …	
… upper-mid level morphotoneme [44] on S1	… lower-mid level morphotoneme [22] on S1
kɔ kʰɔ 高考 33.323 *college entrance exam*	ɔŋ ɕy 洪水 23.323 *flood*
… mid rising morphotoneme [323] on S1	… short mid stopped morphotoneme [3] on S1
ɓɯa ɕie 保險 43.212 *insure*	tɕʰyə̆ kʰəu 出口 4.323 *export*
… high falling morphotoneme [51] on S1	… depressed high falling morphotoneme [241] on S1
ɗe ɓi 對比 33.323 *compare*	z̩ nia 字典 32.212 *dictionary*
… lower-mid rising morphotoneme [212] on S1	… low rising morphotoneme [13] on S1
ɓa kuo 百果 43.212 *all kinds of fruits*	ba kuo 白果 32.212 *ginko*

Table 1 indicates, firstly, five pitch shapes for the first syllable tone: upper-mid level [33], low rising [23], high and mid falling [43], [32], and short stopped high [4]. These shapes reflect several complex mergers. The mid level [33] tone is the realization of a merger between the underlying high falling tone and the underlying upper-mid level tone. The high falling [43] tone is the realization of a merger between mid rising and lower-mid rising tones. The mid falling [32] tone represents a merger between underlying low rising and depressed high falling tones. The low rising [23] tone corresponds to underlying low level tone, and the short high [4] tone corresponds to short stopped mid. Note that none of the resulting first syllable tonal allomorphs corresponds to its citation shape. A high falling citation tone is realized as mid level, for example, and a high rising citation tone is realized as high falling. The same sort of first tone complexity was also demonstrated for the Wu dialect of Wencheng [9], and appears typical.

In contrast to the first syllable tones, the word-final tone is straightforward. Table 1 shows it has two surface forms, both with delayed rising pitch like the corresponding citation tone. One is in the mid pitch range, represented as [323], and one slightly lower "[212]"). The conditioning is very largely clear: the lower version occurs after preceding falling pitched tones, the higher elsewhere. Exactly the same intrinsic perseverative allotony conditioned by a [+/- fall] on the preceding syllable occurs in Wencheng [9], showing that even in right-dominant systems the weak tone can influence the strong. The lower allotone also occurs after the short high tone, which has a falling F0, but is too short to have a pitch contour. The conditioning is not clear in this case.

Figure 3 shows the mean tonal acoustics corresponding to the shapes in table 1 (colour-coding is used for the first-syllable tones). The mid rising citation tone acoustics are also shown. Five different mean F0 shapes – two falling, one level, one rising and one short – can be seen for the first syllable tone corresponding to the five pitch shapes just described. The two dipping F0 shapes corresponding to the two intrinsic word-final allotones can be seen lying a little lower than the citation tone. Figure 3 shows the word-final tone can be considered as a mid rising citation target intrinsically perturbed by co-articulation with the preceding syllable tones: a clear case of word-final preservation of tone.

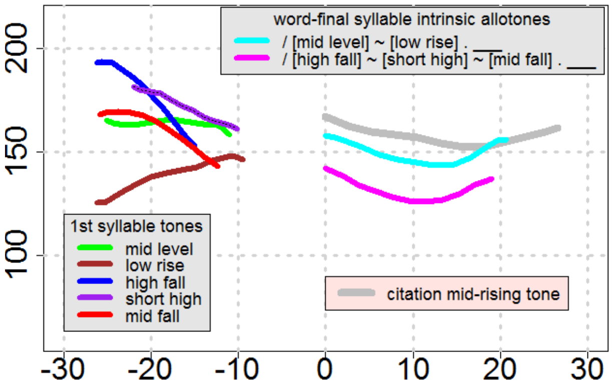


Figure 3. *Mean tonal acoustics of Maodian disyllabic words with underlying mid rising tone on word-final syllable. X-axis = duration (csec.) aligned at onset of second syllable rhyme, y-axis = F0 (Hz).*

6. Disyllabic tone sandhi: a more complex example

Table 2 gives examples of words with underlying lower-mid level [22] tone on the word-final syllable, and the corresponding acoustics are shown in figure 5. The mean acoustics of the low level citation tone have also been plotted. It is evident that the situation is more complex for both first syllable and word-final tones. Unlike before the mid rising tone just discussed there are no first syllable mergers. There are eight different tonal shapes on the first syllable. Their F0 shapes of are all clear in figure 4. Again none match their citation tones. Risking confusion, we list them here (they are colour-coded with table 2 to help matching). As in the previous section, the lower-mid level citation tone corresponds to a low rising tone (brown), the mid rising tone corresponds to a high fall (blue), the high falling citation tone corresponds to mid level (green), and the depressed high fall corresponds to a mid fall (yellow). Unlike the previous section, the upper-mid level citation tone corresponds to a high rising tone (red), the low rise tone corresponds to a low concave tone (magenta), the mid short stopped citation tone corresponds to a short high rise (purple) and the short stopped rising tone corresponds to a low level (orange). None of these look like morphotonemic alternations easily generalizable with conventional tone features.

Table 2. *Examples of Maodian speaker's tone sandhi in words with lower-mid level morphotoneme* [22] *on word-final syllable. Pitch representations are color-coded with figure 4.*

word-final lower-mid level morphotoneme [22] preceded by …	
… upper-mid level morphotoneme [44] on S1	… lower-mid level morphotoneme [22] on S1
$\mathrm{t^h}$ia dʑɔ 天橋 45.41 *flyover*	jɔ mɯa 羊毛 24.51 *wool*
… mid rising morphotoneme [323] on S1	… low rising morphotoneme [13] on S1
ɓɯa dz̩ 保持 53.21 *preserve*	du bi 肚皮 243.31 *stomach*
… high falling morphotoneme [51] on S1	… depressed high falling morphotoneme [241] on S1
$\mathrm{t^h}$a bɐŋ 太平 33.334 *peace*	di dʑiəu 地球 32.22 *globe*
… lower-mid rising morphotoneme [212] on S1	… short stopped low rise morphotoneme [12] on S1
tsǒ dʑiəu 足球 34.51 *soccer*	zɛ̌ dəu 石頭 2.223 *stone*

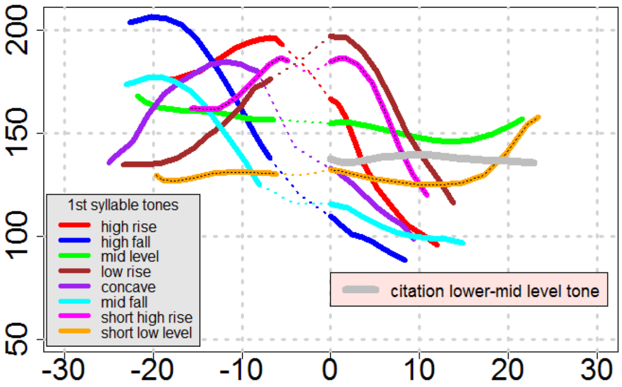


Figure 4. *Mean tonal acoustics of Maodian disyllabic words with underlying lower-mid level tone on word-final syllable. X-axis = duration (csec.) aligned at onset of second syllable rhyme, y-axis = F0 (Hz). Dotted lines indicate F0 on intervocalic consonant.*

The word-final tonal pitch shapes fall into three classes. There are, firstly, four pitches falling from different heights to low: [51], [41], [31] and [21]. The onsets of the five F0 shapes corresponding to these four falling pitches are clearly all determined by the trajectory of the preceding tone, and therefore can be considered intrinsic variants of a falling pitched tone (or even realisations of no tone, with their falling pitch conditioned by a low boundary tone). Secondly, there are two pitches rising to mid: [224] and [334]; these too are intrinsically determined by the height of the preceding tone. Finally, there is one low level [22] pitch, its percept corresponding to the clear leveling out of the F0. There are thus three extrinsic allotones of the word-final lower-mid level tone: falling, rising and level, only the last of which, just, can be considered a case of preservation. Moreover, a case can be made for combinations with the entirely predictable falling word-final allotone to be instantiations of *strong-weak* metrical structure rather than the *weak-strong* structure implied by right dominance. Two of the shapes – 53.21 and 243.31 – are reminiscent of spread high falling and depressed high falling first-syllable tones, except there is no explanation for where such first-syllable tones might have come from: (depressed) high falling and convex tones exist in Maodian to be sure, but table 2 shows they are related to first-syllable [33] and [32] shapes! It seems that, for these data at least, right dominance is not monolithic, and word-final tone preservation cannot be considered criterial for it.

7. Summary

A small part of the tone sandhi behavior has been described for a speaker of Wuzhou Wu, and the data interrogated for typical right-dominant behavior using impressionistic description and quantified acoustics. Complex first syllable behavior typical of right dominant systems was observed. The mid rising tone was shown to be preserved word-finally, but not the lower-mid level tone, showing that preservation of word-final tone is not invariant in putatively right-dominant systems. Clearly, right-dominance is worthy of further study.

8. Acknowledgements

Many thanks to our two anonymous reviewers!

9. References

[1] Ballard, W., "Wu, Min and a Little Hakka – Tone Sandhi: Right and Left", Cahiers de Linguistique Asie Orientale 13:3-34, 1984.

[2] Zhang J., "A directional asymmetry in Chinese tone sandhi systems", J. East Asian Linguistics, 16:259-302, 2007.

[3] Pan W., "An Introduction to the Wu dialects", in Wang S. [Ed.] Languages and Dialects of China, Journal of Chinese Linguistics Monograph 3:237-293, 1991.

[4] Ballard, W., "Oujiang Wu Tone Sandhi: Visi-Pitch Results", Chinese Language and Linguistics 1: Chinese Dialects. Symposium Series of the Institute of History and Philology Academia Sinica 2, Taipei, 41-66, 1992.

[5] Chao Y. 趙元任，現代吳語的研究 *Studies in the Modern Wu Dialects*, Tsing Hua College Research Institute Mono. 4, 1928.

[6] Qian N. 钱乃荣，当代吴语研究 [*Studies in the Contemporary Wu Dialects*], Shanghai Educational Press, 1992.

[7] Zhang J., A Sociophonetic Study on Tonal Variation of the Wúxī and Shànghǎi Dialects, LOT Netherlands Graduate School of Linguistics, 2014.

[8] Fu G. 傅国通, Fang S. 方松熹, Cai Y. 蔡勇飞, Bao S. 鲍士杰, Fu Z. 傅佐之, 浙江吴语分区 [Wu dialect subgroups of Zhejiang], Zhejiang Linguistics Society, 1985.

[9] Rose, P., "Complexities of Tonal Realisation in a Right-Dominant Chinese Wu dialect – Disyllabic Tone Sandhi in a Speaker from Wencheng", Journal of the South East Asian Linguistics Society 9:48-80, 2016.

Towards a Better Understanding of Regional Variation in Standard Australian English: Analysis and Comparison of Tasmanian English Monophthongs

Rael Stanley

University of Melbourne

raels@student.unimelb.edu.au

Abstract

Using phonetic analysis, this investigation looks at the acoustic properties of Tasmanian English vowels, as produced by speakers of that variety of English from the Austalk corpus. It compares the formant values of monophthongal vowel targets to published formant values Melbourne and Sydney vowels. The aim of the study is to give a first outline of the vowel space of Tasmanian English, to determine whether there is any regional variation between Tasmanian and mainland vowel realisation, and to compare what differences there are in vowel realisations for older and younger speakers of Tasmanian English.

Index Terms: Regional variation, Tasmanian English, vowels

1. Introduction

This study is looking at the accent produced by speakers of Tasmanian English, in comparison to varieties of Australian English spoken in other Australian states. As vowels are the phonemes most responsible for accent variation [1], it will be focussing on how production of them is similar to or different from vowels produced elsewhere.

1.1 Research Questions

The bulk of work done on regional accent variation for Australian English has focussed on the larger population centres in the country, such as Sydney, Melbourne, Adelaide, and Perth. However, there is a paucity of data for smaller areas, particularly the capital of Tasmania: Hobart.

Separated from the mainland of Australia by the waters of Bass Strait, Tasmania is a mountainous island that was, for much of its history post-British-colonisation, isolated from the rest of Australia's population by more than simply the great distances found between populations centres on the mainland but also the broad and treacherous sea.

As physically isolating barriers are a common factor in regional variation of languages and dialects [2], it is somewhat surprising that there has been so little research into differences between Tasmanian English and the other varieties spoken in Australia, and this study seeks to redress this, with the following research questions:

1. What are the acoustic properties of Australian English short and long monophthongs as spoken by Tasmanians?
2. Are there differences in these acoustic properties that are present according to age categories?
3. Is there regional variation present between the vowels analysed in this study, and those for Melbourne and Sydney, in studies by Cox and Billington [3] & [4] respectively?

2. Method

2.1 Data Collection and Participants

21 male and 18 female speakers of Hobart English produced vowels in citation form (/hVd/), and were extracted from the list of Tasmanian speakers, who had completed all their schooling in Tasmania, from the AusTalk corpus [5]. A total of 1,096 vowel tokens were downloaded for analysis via Alveo online [6].

The speech data for each group, male and female, was separated into younger and older speakers, with younger speakers being aged between 20 and 39 years, and the older speakers aged 60 years and over, excluding those speakers aged 40 to 59 years in order to more clearly see what effects age have on the Tasmanian English accent. Table 1, below, shows the distribution of speakers across age categories and gender.

Table 1. *Showing number of speakers of each gender, according to age category*

Gender	Number of Younger Speakers	Number of Older Speakers
Female	11	7
Male	13	8

2.2 Data Labelling

The vowels chosen for analysis were the short and long monophthongs of Australian English, because these are the vowels shared by the analyses Performed by Cox and Billington [3] & [4] that I am comparing my data with.

Segments for each repetition of each cited form were automatically labelled through use of the application WebMAUS Basic [7].

2.3 Analysis

The open-source data manipulation program RStudio [8] was used to automatically identify the formant values for F1 and F2 at the vowel midpoints and map them to ellipse plots for comparison between the different groups by both age and gender of the speakers. Any outliers in the data sets were identified visually from these plots, and had their formants manually checked and adjusted, where necessary, as suggested by Harrington [9] using Praat [10].

Linear Mixed Model (LMM) tests were run, using the packages lme4 [11] and multcomp [12] to provide statistical significance data on the differences in these values. The LMM tests were run on both F1 and F2 values, using age category and gender as fixed effects, with speaker as a random effect for both.

Mean formant data collected for each vowel for male and female groups of the younger speakers was compared to mean formant data for the same monophthongal vowels collected in [3] and [4].

3. Results

Running LMM tests for age category as a fixed effect showed more sporadic statistical significance across the mean F1 and F2 data for males and females. The only vowels that showed statistically significant differences in both F1 and F2 by age category were /æ/ for both male and female speakers, and /eː/ for female speakers, only. Those vowels which displayed statistical significance in mean F1 value across age categories were female speakers' productions of /e, ɜː, oː/ and male speakers' productions of /eː, e, ɜː/, while the vowels showing differences of statistical significance in mean F2 values were female speakers' productions of /ɐː, ɔ, ʊ, ʉː/ and male speakers' productions of /ɐː, ɐ, ʉː/.

3.1 The Short Vowels of Tasmanian English

A pair of ellipse plots of the short vowels produced by the younger and older groups of Tasmanian females in citation forms is presented in Figures 1 and 2, below. These plots have been chosen, as they most clearly display the points of interest in these data sets.

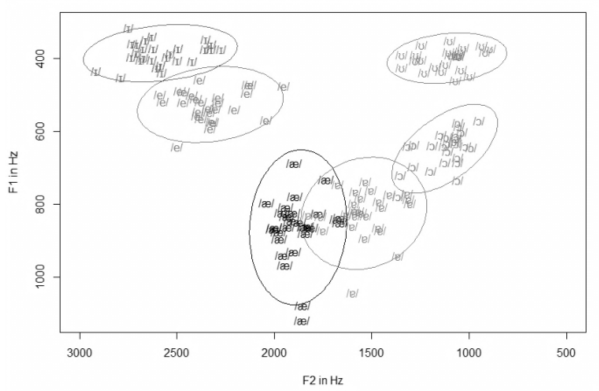


Figure 1: *Ellipse plots of the F1/F2 values of short vowels produced by younger female speakers of Tasmanian English*

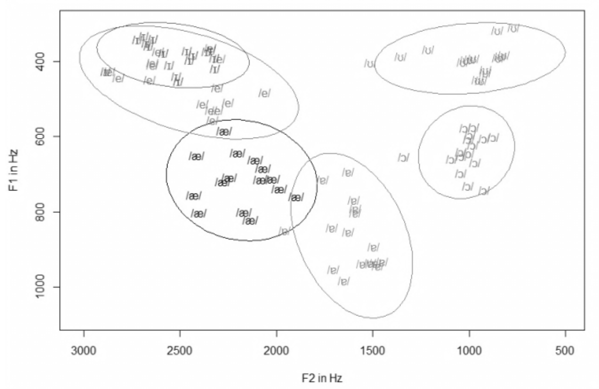


Figure 2: *Ellipse plots of the F1/F2 values of short vowels produced by older female speakers of Tasmanian English*

As can be seen in the above pairs of plots, there is a notable retraction and lowering of the /æ/ ("had" vowel) by the group of younger speakers, as compared to the older group. The mean values for F1 and F2 in the younger group are 863Hz and 1878Hz, respectively, while the same mean values for the older group are 716Hz and 2179Hz. The higher F1 (i.e. a more open vowel) values for younger speakers – a difference of 147Hz (p ≤ 0.001) – between the two groups of speakers patterns with findings in [13], showing that short front vowels appear to be undergoing a reversal of the sound change that for a time saw the vowels raising. Further, the retraction is clear – a difference of 301Hz (p ≤ 0.001) in the mean F2 values of the groups – also patterning with the findings [13].

Another point to note is the comparatively large overlaps shown by the ellipses for the "head" and "hid" vowels, /e/ and /ɪ/ respectively, in older speakers. The ellipse for production of /ɪ/ is almost entirely covered by the ellipse for production of /e/. However, for individual speakers there is a clear difference between vowel height in production of /e/ and /ɪ/ - /ɪ/ vowels for a speaker have a mean F1 value that's 73Hz lower than the corresponding speaker's /e/ vowel (p ≤ 0.001).

These same patterns are seen when comparing the data for older and younger groups of male speakers

Overall, it can be seen that both genders of younger speakers appear to be reversing previously found raising and fronting of /æ/, patterning with the data found in [13]. While older speakers of both genders have /e/ and /ɪ/ productions that are very closely clustered (the females more so than the males), the individual speaker productions of these vowels are distinct from one another. And, while younger female speakers have considerably more distinct ellipse plots than their older counterparts, this is only true with regards the high front vowels for male speakers (who also a very broad range of realisations for /æ/, not seen in older speakers).

3.2 Frequency Data for the Short and Long Monophthongs of Tasmanian English for Younger Speakers

In this section, a description of the vowels of Tasmanian English is put forward. This incorporates the data shown in the previous sections, as well as the below figure (3) showing the mapped vowel spaces for both male and female younger speakers.

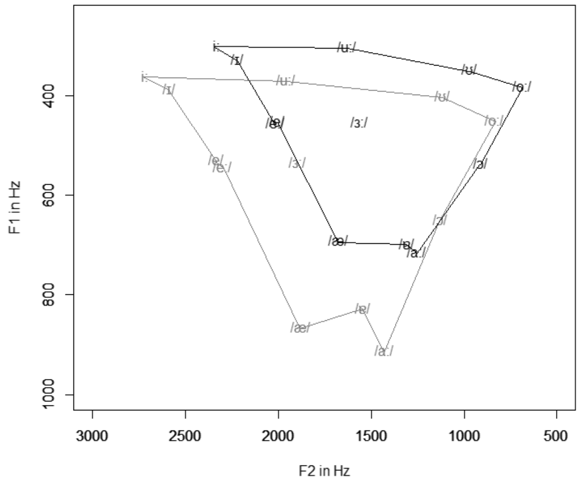


Figure 3: *A plot of the mean F1 and F2 midpoint values for the long and short vowels produced by younger female (black) and male (grey) speakers from Hobart.*

The most readily obvious gender difference seen in this figure is the difference in relative position for the low vowels: /ɐ:/ and /ɐ/ are where we see the first big differences between the two data sets. For male speakers, both these vowels are sitting very close together, with the short vowel only slightly higher and more forward than the long vowel. However, for the female speakers, the short vowel /ɐ/ is considerably higher than the long vowel /ɐ:/, being even higher than /æ/. In this way, the short vowel /ɐ/, for female speakers patterns similarly to how it does for Melbourne data in [4], while the long vowel /ɐ:/ is lower than /æ/, putting it in a relative position more similar to the Sydney data in [3].

3.3 Regional Variation

Below, in Figures 4 and 5, can be seen the comparative vowel spaces for female speakers from Hobart, Melbourne, and Sydney. Overall, the data for female speakers from Hobart have a general vowel space that appears retracted and raised, compared to the speakers from Melbourne and Sydney.

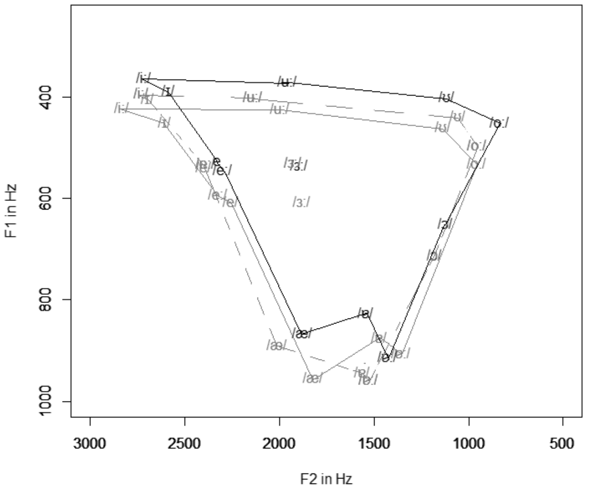


Figure 4: *Vowel spaces for Hobart female speakers (in black), Melbourne female speakers (in grey), and Sydney female speakers (the broken line)*

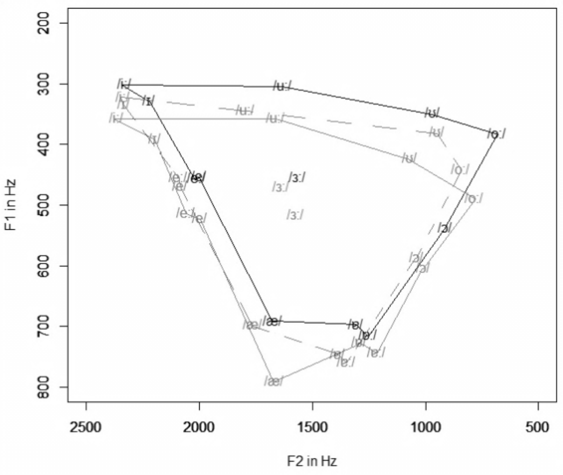


Figure 5: *Vowel spaces for Hobart male speakers (in black), Melbourne female speakers (in grey), and Sydney female speakers (the broken line)*

3.3.1 "Had" /æ/

The open front vowel /æ/ shows quite noticeable variation between the three states. The difference between Hobart and Melbourne is greater than that between Hobart and Sydney. Compared to speakers from Sydney, the speakers from Hobart show retraction and very slight raising in this vowel, with an F1 frequency difference of 23Hz and an F2 frequency difference of 135Hz. Compared to Melbourne data, /æ/ produced by speakers from Hobart is higher in the vowel space, the difference in F1 frequency between the two being 88Hz.

3.3.2"Hard" /ɐ:/ and "Hud" /ɐ/

These two vowels show a very similar pattern in their F1 and F2 frequencies between Hobart and Melbourne data sets. Both these vowels are almost plotted on top of one another for the speakers from Sydney. However, there is a clear difference in both the Hobart and Melbourne productions of these vowels, with the short vowel /ɐ/ being higher and further forward in the vowel space than the long vowel /ɐ:/. Both Hobart and Melbourne-based speakers have very similar targets for /ɐ:/ but Hobart speakers appear to produce an /ɐ/ that is higher and slightly further forward in the vowel space than speakers from Melbourne. The difference in F1 values for this vowel is 48Hz, and the difference in F2 values is 72Hz.

3.3.3 "Who'd" /ʉ/

Realisations of this vowel share a similar relationship between speakers from Hobart and speakers from Melbourne, as the vowel in "Hood". The difference isn't particularly great but the Hobart data set shows a lower mean F1 value for speakers from Hobart than is seen for speakers from Melbourne. The Sydney data, however sits roughly at the midpoint between Hobart and Melbourne, in F1 frequency value but is further forward in the vowel space, with a higher F2 value than either of the others. This pattern is repeated for male speakers.

4. Discussion

This study analysed variability in Standard Australian English spoken in Tasmania. In bringing Tasmania into the analysis, there were four research questions posed. This section examines the findings surrounding each of those questions.

4.1 A Description of the Vowel Space of Hobart English

The results of the acoustic analysis of the vowels of English as spoken in Hobart, showed vowel spaces shaped rather typically for Australian English, although there was some variability for some sounds. For example he "Had" vowel /æ/ is less open than that produced in Sydney or Melbourne (in particular), however it still shows patterning similar to that found in [13], in reversing the raising and fronting movement previously observed. This is a trend that is clearer in the F1/F2 frequency means for female speakers than it is for the male speakers, a point that is to be expected, considering that females have a tendency to lead males with regards to linguistic changes [14]. There is also a distinctly higher production of /ɐ/ as compared to /æ/ for female speakers than there is for male speakers.

4.2 Difference in Age Category

As discussed throughout the results, age differences occur for many of the vowels, indicating that the accent changes observed in [13] are, indeed, also occurring in a similar manner in Tasmania as in other regions. Most vowels show differences in the mean formant values for one or both of F1 and F2 of their targets between younger and older groups, for both males and females. Overall, where the values show statistically significant differences, the younger speakers have higher mean formant frequency values than the older speakers, resulting in a lower and more front set of vowel realisations.

4.3 Regional Variations

There is some visible variation in the overall vowel plots between Tasmania, Victoria, and New South Wales, for both males and females, with what appears to be a general trend of a compressed vowel space for Tasmanian speakers, compared to their mainland counterparts. While this is not very great in extent, there are some larger differences to be seen in some of the low vowel targets, specifically the realisations of /æ/ show large variance between Tasmanian and Victorian female speakers of both genders, and there is a similarly large gap between production of /ɐ/between female Tasmanians and females from New South wales. There is also a clear difference in the data for the production of /o:/ between younger speakers from Tasmania and Victoria, with that vowel being produced noticeably higher and further back in the vowel space of Tasmanians, and speakers from New South Wales occupying a space in between the two. Conversely, the Tasmanian production of /ʉ:/ patterns similarly to Victorian production. That is, it's still quite a fronted vowel but not nearly to the same extent as that seen in New South Wales.

5. Conclusion

This study has added to the current knowledge on variation in Standard Australian English, by adding acoustic data for the vowels of Tasmanian English to the discussion, displaying vowel spaces for both male and female speakers. The study has shown evidence for some regional variation between reports on Standard Australian English as spoken in Sydney and Melbourne in some vowels, as well as variation between older speakers of this variety of Standard Australian English.

Being that this study was a one of the monophthongs, it would make sense that further study could be made into how the diphthongs of Tasmanian English pattern. Also, since vowel targets were analysed at a fixed point, a dynamic analysis of the vowel formants would likely be enlightening on this topic, in future.

6. References

[1] F. Cox, Australian English Pronunciation and Transcription, Melbourne: Cambridge University Press, 2012.

[2] R. Wardhaugh, An Introduction to Sociolinguistics, 5th ed., Carlton: Blackwell Publishing, 2006.

[3] F. Cox, "The Acoustic Characteristics of /hVd/ Vowels in the Speech of Some Australian Teenagers," *Journal of Australian Linguistics,* pp. 147-179, 2006.

[4] R. Billington, "Location, Location, Location! Regional Characteristics and National Patterns of Change in the Vowels of Melbourne Adolescents," *Australian Journal of Linguistics,* pp. 275-303, 2011.

[5] D. Burnham, D. Estival, S. Fazio, J. Viethen, J. Cox, R. Dale, S. Cassidy, J. Epps, R. Togneri, M. Wagner, Y. Kinoshita, R. Göcke, J. Arciuli, M. Onslow, T. Lewis, A. Butcher and J. Hajek, "Building an audio-visual corpus of Australian English: large corpus collection with an economical portable and replicable Black Box," in *Proceedings of the 12th Annual Conference o the International Speech Communication Association (Interspech, 2011)*, 2011.

[6] S. Cassidy, D. Estival, T. Jones, D. Burnham and J. Berghold, "The Alveo Virtual Laboratory: A Web Based Repository API," in *9th Language resources and Evaluation Conference (LREC 2014)*, Reykjavik, 2014.

[7] T. Kisler, F. Schiel and H. Sloetjes, "Signal processing via web services: the use case WebMAUS," in *Proceedings Digital Humanities*, Hamburg, Germany, 2012.

[8] RStudio Team, "RStudio: Integrated Development for R.," Boston, 2015.

[9] J. Harrington, The Phonetic Analysis of Spech Corpora, Munich: Blackwell Publishing, 2010.

[10] P. Boersma and D. Weenink, "Praat: doing phonetics by computer [Computer program]. Version 6.0.17," 2016. [Online]. Available: http://www.praat.org. [Accessed 21 April 2016].

[11] D. Bates, M. Mächler, B. Bolker and S. Walker, "Fitting Linear Mixed-Efects Models using lme4," *Journal of Statistical Software,* vol. 67, no. 1, pp. 1-48, 2015.

[12] T. Hothorn, F. Bretz and P. Westfall, "Simultaneous Inference in General Parametric Models," *Biometrical Journal,* vol. 50, no. 3, pp. 346-363, 2008.

[13] F. Cox and S. Palethorpe, "Reversal of short front vowel raising in Australian English," in *Proceedings of Interspeech 2008, 22nd-26th September 2008*, Brisbane, 2008.

[14] W. Labov, Principles of Linguistic Change, Oxford: Blackwell, 1992.

Background Specificity in Forensic Voice Comparison and Its Relation to the Bayesian Prior Probability

Michael Wagner[1], Yuko Kinoshita[2]

[1]Faculty of ESTeM, University of Canberra
[1]Research School of Computer Science, The Australian National University
[1]Quality and Usability Lab, Technical University of Berlin
[2]College of Arts and Social Science, The Australian National University
michael.wagner@canberra.edu.au; yuko.kinoshita@anu.edu.au

Abstract

This study investigates the effect of background data specificity on likelihood ratio and prior odds, and consequently on the posterior odds outcome. It is motivated by discussions on the correct choice of speaker recognition background, particularly in forensic voice comparison. We performed strictly controlled experiments with the ANDOSL database where background specificity is the sole independent variable. Results show that target and non-target scores are better separated with less specific background, but that in turn priors must be adjusted down. Because the risk of class recognition instead of individual recognition increases with lower background specificity, we suggest that the prior probability in the Bayes formula is factorised into one part that remains in the domain of the trier of fact – as is conventional – and another part that is related to the specificity of the assumed or agreed background.

Index Terms: Forensic voice comparison, Bayesian method, forensic prior probability, background specificity.

1. Introduction

The Bayesian approach used in forensic voice comparison (FVC) is similar in principle to that of non-forensic speaker authentication tasks. However, the interpretation of Bayesian likelihood ratios (LRs) is quite different in the FVC context.

In non-forensic speaker authentication, the Universal Background Model (UBM) has long been the standard method [1, 2, 3, 4]; contemporary methods, such as joint factor analysis, iVectors etc. are also implicitly based on the chosen UBM. Systems such as those prominent in recent NIST speaker recognition evaluations [5], generally use very large UBMs that represent speaker characteristics across dialects, accents and even languages spoken in multicultural societies, except that by common consensus, they are usually restricted to speakers of the same sex as that of the unknown speaker. To compensate for external factors, such as environmental noise and channel characteristics, speaker recognition scores are normalised with a cohort of speakers with similar attributes.

In FVC, the situation is somewhat more complicated. Within the context of the widely accepted Bayesian paradigm [6, 7], the forensic speech scientist estimates the likelihood ratio (LR) between the two likelihoods: 1) for the crime-scene recording to be consistent with the speaker model of the suspect (numerator of the LR) and 2) for the same recording to be consistent with the multi-speaker model of a background population (denominator of the LR). Although it is rarely performed explicitly in reaching the final decision, the trier of fact is required to combine the LR obtained from FVC with the prior odds $P(H_{so})/P(H_{do})$: the prior probability of the same-origin hypothesis (H_{so}–offender and suspect are the same person) versus the prior probability of the different-origin hypothesis (H_{do}–offender and suspect are different persons).

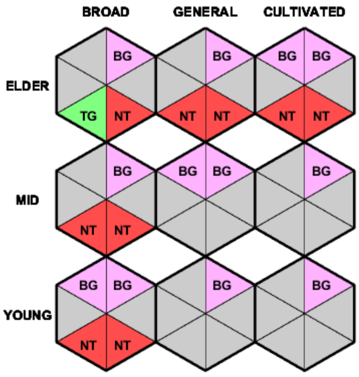


Figure 1. *Partitioning of the non-accented male speakers of ANDOSL into 9 subgroups, each of 6 speakers, and speaker partitioning into target (TG), non-target (NT) and background (BG) speakers.*

The determination of the prior odds is usually considered the domain of the court, as the forensic scientist does not have access to information on the case other than the voice recordings. Combining those prior odds with the forensic scientist's LR $p(X|H_{so})/p(X|H_{do})$ yields the posterior odds $P(H_{so}|X)/P(H_{do}|X)$ for the same-origin hypothesis versus the different-origin hypothesis according to the Bayes Rule [7]

$$\frac{P(H_{so}|X)}{P(H_{do}|X)} = \frac{P(H_{so})}{P(H_{do})}\frac{p(X|H_{so})}{p(X|H_{do})}. \qquad (1)$$

In FVC, H_{do} plays a pivotal role in selecting the background population, and this has been the subject of much debate among forensic scientists. Some have argued that the background population should be tailored to the characteristics of the offender, the suspect or both [8]. It has also been argued that the background population should be based on those characteristics of the offender's voice that both prosecution and defence agree upon [6] or that it should represent a set of speakers sufficiently similar to the offender's voice that an investigating police officer would bother submitting voice samples for examination by a forensic scientist at all [9].

In forensic casework, the different-origin hypothesis defines the subpopulation to which the offender apparently belongs and to which any suspects should also belong. Those attributes of the subpopulation that can affect speech acoustics (e.g. language variety, gender, age range) will set the selection criteria for the background population data for the case. As should be clear from Eq. (1), the selection of the background affects the forensic scientist's LR through $p(X|H_{do})$ as well as the trier of fact's assignment of prior odds through $P(H_{do})$.

A commonly used imaginary forensic examination illustrates the two effects: Assuming a criminal investigation on an island of 100 inhabitants, if nothing other than being a member of the island population is known about the offender, the prior odds $P(H_{so})/P(H_{do})$ would be 1/99, and the forensic scientist should build a background model from a representative sample of the entire island population. If however, in addition, the offender were known to be a member of the female half of the population, the prior odds would increase to 1/49, and the forensic scientist would build the background model only from the female subpopulation. Any additional knowledge about the offender would further raise the prior odds and, at the same time, likely diminish the acoustic variance of the background.

While the characteristics of the background population such as gender or dialect should be consistent with an agreed H_{do}, in practice the forensic scientist's choices are often constrained by data availability, time and resources. Sometimes there is not even a clear reference to an agreed H_{do}. This is problematic for the validity of the Bayesian estimation of the posterior odds unless the scientist informs the trier of fact of this relationship between H_{do} and choice of background data on one hand, and LR, prior and posterior odds on the other.

In the remainder of this paper, we thus examine how the specificity of the different-origin hypothesis and the corresponding selection of the background population affect the outcome of forensic voice comparison: firstly through $p(X|H_{do})$ and the resulting LR and secondly through $P(H_{do})$ and its effect on the prior odds estimated by the trier of fact.

2. Experiment

2.1. Data

The Australian National Database for Spoken Language (ANDOSL) [10] comprises Australian English speech data from a range of speakers, varying in age, sex, and their variety of Australian English. For this study, we utilise only the read-sentence data by the ANDOSL male native speakers. Within that population, we have a 3×3×6 partitioning into 3 age groups, elder, mid, young, the 3 sociolect groups of Australian English, broad, general, cultivated, on the basis of the tag provided in ANDOSL [11], and 6 speakers in each group, as is illustrated in Fig. 1.

Each of the 9 subpopulations is shown as a hexagon, and the 6 speakers of each subpopulation are shown as colour-coded triangles. The single target speaker is shown as the green triangle, tagged *TG*. The non-target speakers are shown as red triangles, tagged *NT*, 5 in the first row for the *elder* subpopulation and 5 in the first column for the *broad* subpopulation. The 12 background speakers, tagged *BG*, are shown as magenta triangles. This design ensures that there is no overlap between target, non-target and background speakers, hence avoiding a potential statistical bias.

Each of the 54 male speakers read 200 sentences that were designed to cover the entire acoustic-phonetic space of Australian English. Of those, 180 were used solely for training background models. Using 180 sentences for UBM training, we assume that the background models cover the acoustic-phonetic space of the background population sufficiently. Of the remaining 20 sentences, 10 were used solely for maximum-a-posteriori (MAP) adaptation of the target-speaker models, and 10 were used solely for the target and non-target testing. We consider using 10 sentences for GMM adaptation forensically realistic, given the typical constraints of FVC casework, where suspects are often uncommunicative during police interviews and provide precious little material for the adaptation of the target-speaker GMM.

Recordings of the 3×3×6×200=10,800 sentences are stored as wav files, sampled at 20,000 samples/s and 16 bits/sample. 12 mel-frequency cepstral coefficients (MFCC) and log energy were determined for 20ms windows shifted in 10ms steps. Derivative coefficients were discarded as the amount of data was insufficient for training higher-dimensional models. Using a simple absolute energy threshold, low-energy frames were eliminated and about 61% of the frames retained, yielding on average 313 feature vectors per sentence for the analysis.

2.2. Experimental design

The read-speech data in ANDOSL were produced under highly controlled conditions, and each speaker was recorded in a single session. ANDOSL is thus generally regarded as an inadequate database for FVC experiments. However, our experimental design turns this limitation into an advantage. The single-session nature of ANDOSL and the read-speech material enabled perfect control over the independent variables of our design. Being a single session recording eliminates extraneous variation such as intersession and channel variation as well as intra-speaker variation in health or emotion. Using speech material read from prepared texts, we exclude the variability in quantity and phonetic contents that is inevitable with spontaneous speech data. Therefore, we can reasonably interpret any effects on the output as being caused by the chosen background specificity–the independent variable of our design.

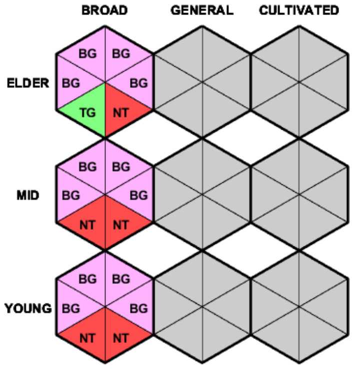


Figure 2. *Speaker verification in the broad subpopulation with specific background.*

We conducted altogether 4 experiments. In the first, the target speaker, the 5 non-target speakers and the 12 background speakers are all from the subpopulation of the *broad* sociolect speakers as shown in Fig. 2. In the second experiment, the target speaker, the 5 non-target speakers and the 12 background speakers are all from the *elder* subpopulation as shown in Fig. 3. Each experiment proceeded to build a UBM from the background speakers, building a GMM for the target speaker, and determining LRs for target and non-target tests.

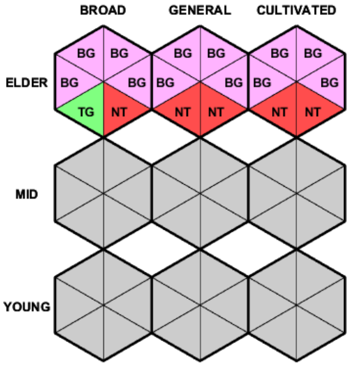


Figure 3. *Speaker verification in the elder subpopulation with specific background.*

We then repeated these 2 experiments altering the specificity of the UBM by building it from the background speakers drawn from the entire population of male speakers as shown in Fig. 1, while the other factors, i.e. target and non-target speakers and the sentence material, were kept identical. For the comparison of system performance, LLRs are usually normalised or calibrated for environmental or channel variation between recordings. However, the current study does not require such a step due to the strict control in experimental design.

Fig. 4 depicts our experimental design schematically. We constructed a UBM from the 180 designated sentences spoken by the 12 chosen background speakers. For the target speaker, we MAP-adapted this UBM to a target-speaker-specific GMM using the 10 designated adaptation sentences of the target speaker. Finally, 10 target trials were conducted with the designated test sentences of the target speaker, and 50 non-target trials were conducted with the same test sentences spoken by 5 non-target speakers. For each trial, we produced the sentence-mean log likelihoods and log likelihood-ratios with respect to the target GMM and the UBM.

For the *broad*-sociolect subpopulation, we conducted the following pair of experiments: firstly, we built a UBM from the 12 designated background speakers of the subpopulation as illustrated in Fig. 2. For the designated target speaker, we adapted the UBM to build the target-speaker GMM and conducted the target trials. Then we conducted the non-target trials with the designated 5 non-target speakers of the subpopulation. Both target and non-target trials consist of the designated 10 testing sentences for each target and non-target speaker. That experiment was repeated with the same target and non-target speakers, but with the 12 background speakers drawn from the full population as shown in Fig. 1 and the target-speaker GMMs adapted from that wider-background UBM.

An equivalent set of experiments was then conducted with the other subpopulation under investigation, *elder* speakers. Here, the background speakers were drawn from this subpopulation and from the full population as shown in Fig. 3 and again in Fig. 1. For both pairs of experiments, the independent variable is the specificity of the background: either specific to the subpopulation matching the test speaker characteristics or less specific by being a superset of that subpopulation–in our case the entire population of the male speakers in ANDOSL.

For each trial, we observe the log likelihoods (LL) $log\,p(X|H_{so})$ and $log\,p(X|H_{do})$ for each sentence X, each being the mean of the frame log likelihoods for the sentence. We also observe the resulting log likelihood-ratios $LLR = log\,p(X|H_{so}) - log\,p(X|H_{do})$. The statistics of the above LLs and LLRs are the dependent variables of the design, while background specificity is the independent variable.

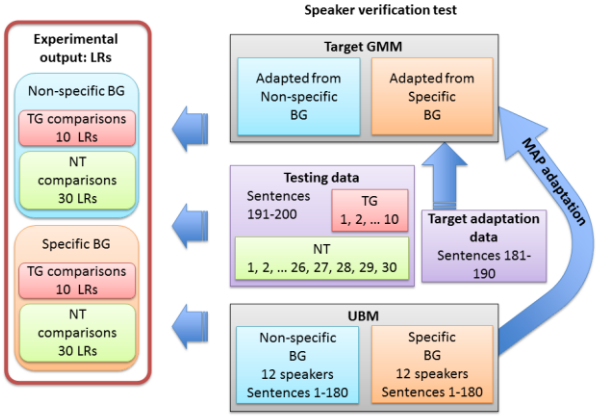


Figure 4. *Experimental speaker recognition system showing the parallel evaluations of non-specific background (blue boxes) and specific background (orange boxes)*

3. Results and discussion

Table 1 presents the mean LLRs for the target and non-target tests for the 2 sets of experiments and their mean differences ΔLLR as well as their log-likelihood-ratio costs C_{llr} [12]. The last 2 rows combine the 2 specific-background and the 2 non-specific-background experiments.

Table 1. *Target and non-target LLRs and their differences.*

Target/UBM	TG LLR	NT LLR	ΔLLR	Cllr
Broad/Broad	2.664	-0.174	2.837	0.496
Broad/Non-specific	2.921	-0.174	3.096	0.491
Elder/Elder	2.682	-0.409	3.091	0.420
Elder/Non-specific	2.921	-0.311	3.232	0.438
Mean Specific	2.673	-0.291	2.964	0.458
Mean Non-specific	2.921	-0.243	3.164	0.464

The results show that the target and non-target scores are separated better for the non-specific background than for the specific background. Fig. 5 shows in addition the distribution of the numerator LLs (LLG) and the denominator LLs (LLU) in Eq. (1). The curves are based on the means and variances of the LLs and a normality assumption for the distributions.

The 2 largely overlapping narrow (green) Gaussians at the right of Fig. 5a show the distributions of the numerator LLs of the *broad* target trials against the specific UBM (dashed line) and against the non-specific UBM (full line). The specificity of the UBM seems to affect neither the mean nor the variance of those LLs. The other 2 narrow (black) Gaussians near the centre of Fig. 5a show the distributions of the denominator LLs of the *broad* target trials against the specific and non-specific UBMs. Those distributions show that the non-specific UBM produces distinctly smaller denominator LLs than the specific UBM. Since the numerator LLs are distributed almost identically, it follows that the non-specific UBM produces higher LLRs than the specific UBM.

Fig. 5b shows the corresponding 4 curves for numerator and denominator LLs against specific and non-specific UBMs for the *elder* subpopulation with the same trends as found for Fig. 5a. Each of the 2 figures also shows 4 wide Gaussians for the non-target trials with the respective specific numerator (green) and denominator (black) LLs closely overlapping and the specific UBM (dashed) yielding a slightly larger mean LL than the non-specific UBM (full) as could be expected.

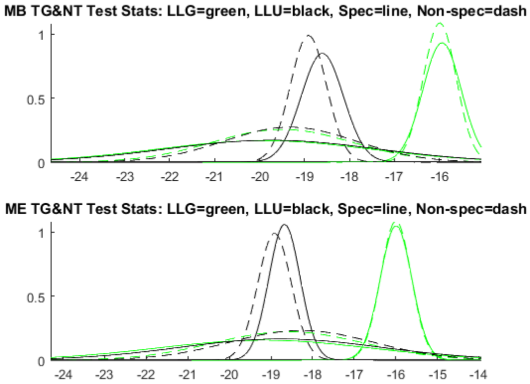


Figure 5. *LL statistics of target trials (narrow) and non-target trials (wide) against target GMM (green) and UBM (black): (a) Broad subset; (b) Elder subset (Note the almost complete overlap between the 4 pairs of wide curves in the 2 figures)*

In both cases, the background specificity affected non-target comparisons less. Our results show a larger distance between target and non-target scores for the non-specific UBM, which seems to be due to the larger denominator LLs of the non-target scores for the non-specific UBM. Table 1 shows no significant differences of C_{llr} between the specific and the non-specific background in either experiment.

From this rather small and therefore limited study, it appears that the specificity of the background is not of major consequence in FVC or even that a less specific background may be preferable for reasons of the slightly larger distance between target and non-target scores found here. However, this interpretation must be weighed against 2 other factors: Firstly, a less specific UBM has the tendency to turn the speaker recognition problem into a recognition of the subpopulation. In other words, there is a danger of recognising, for example, the language variety of a speaker instead of recognising the individual. And secondly, a less specific UBM implies a proportionately smaller prior probability for the defence hypothesis and correspondingly a less conclusive outcome in terms of the posterior odds of the analysis.

For example: an offender is only known to be a member of a population of 8 million Australian adult males. Non-acoustic evidence such as height, eye and hair colour place him in 10% of that population. Using Row 2 of Table 1, the forensic scientist reports a likelihood ratio of $e^{3.096} = 22.109$ against that background population with corresponding posterior odds of $1/7{,}999{,}999 \times 10\% \times 22.109 = 0.276 \times 10^{-6}$.

However, if on the acoustic evidence the forensic scientist determines that the offender is a member of the *broad* accent group of 2 million males and, according to Row 1 of Table 1, reports a likelihood ratio of $e^{2.837} = 17.064$ against that more specific background population, the corresponding posterior odds are $1/1{,}999{,}999 \times 10\% \times 17.064 = 0.853 \times 10^{-6}$, a value about 3 times larger than with non-specific background.

This example illustrates the case for factoring the prior odds into one part that represents the non-acoustic evidence (10% in our example) and another part that represents the size of the background population used by the forensic scientist.

4. Conclusions

A small-scale preliminary study was conducted to investigate how the specificity of the background population affects FVC, using age and sociolect specific subpopulations in the ANDOSL database. LLRs as well as the constituting numerator and denominator LLs were determined dependent on background specificity. Our small-scale experiments show that the denominator LLs for the target speaker were smaller for less specific UBMs and hence those LLRs were larger and the target-non-target separation was larger for less specific UBMs. However, a less specific background bears the risk of inadvertently performing class recognition instead of individual recognition. Further experiments with larger datasets and varying degree of specificity should be conducted.

Also, perhaps more importantly, we must be mindful that the choice of background population directly affects the determination of the prior odds and thus the interpretation of the forensic voice comparison by the trier of fact. In the example presented in this study, assuming equal distribution of the subpopulations, the choice of the full male population of ANDOSL for the UBM would imply prior odds 3 times smaller than the choice of the specific subpopulations.

It is therefore most important for the forensic scientist to report as precisely as possible the characteristics of the background database and its implications for the determination of the LRs and posterior odds of the forensic voice comparison.

5. References

[1] A.E. Rosenberg, J. DeLong, C.H. Lee, B.H. Juang, F.K. Soong, “The use of cohort normalized scores for speaker verification”, *Proc. Int. Conf. on Spoken Language Processing*, 599-602, 1992.

[2] J.B. Millar, F. Chen, M. Wagner, X. Zhu, “The efficacy of cohort normalisation in a speaker verification task under different types of speech signal variance”, *Proc. Austr. Int. Conf. on Speech Science and Technology*, 850-855, 1994.

[3] S. Furui, “Recent advances in speaker recognition”, *Proc. 1st Int. Conf. on audio- and video-based biometric person authentication*, 237-252, 1997.

[4] D.A. Reynolds, T.F. Quatieri, R.B. Dunn, “Speaker verification using adapted Gaussian mixture models”, *Digital Signal Processing*, 10, 19-41, 2000.

[5] G.R. Doddington, M.A. Przybocki, A.F. Martin, D.A. Reynolds, “NIST speaker recognition evaluation - overview, methodology, systems, results, perspective”, *Speech Communication*, 31, 225-254, 2000.

[6] P. Rose, *Forensic speaker identification*, London: Taylor and Francis, 2002.

[7] G.S. Morrison, “Forensic voice comparison” in I. Freckelton and H. Selby [eds], *Expert Evidence*, Ch. 99, Sydney: Thomson Reuters, 2010.

[8] N. Brümmer, E. de Villiers, “What is the ‘relevant population’ in Bayesian forensic inference?”, downloaded on 30 March 2016 from http://arxiv.org/pdf/1403.6008v1.pdf, 2014.

[9] G.S. Morrison, F. Ochoa, T. Thiruvaran, “Database selection for forensic voice comparison”, *Proc. Odyssey 2012*, 62-77, 2012.

[10] J. Vonwiller, I. Rogers, C. Cleirigh, and W. Lewis, "Speaker and material selection for the Australian national database of spoken language", *Journal of Quantitative Linguistics*, 2, 177-211, 1995.

[11] J. Harrington, F. Cox, and Z. Evans, "An acoustic phonetic study of broad, general, and cultivated Australian English vowels," *Australian Journal of Linguistics*, 17:2, pp. 155-184, 1997.

[12] N. Brümmer, J. du Preez, 2006. “Application-independent evaluation of speaker detection,” *Computer Speech & Language*, 20, 230–275.

Author Index