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

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.

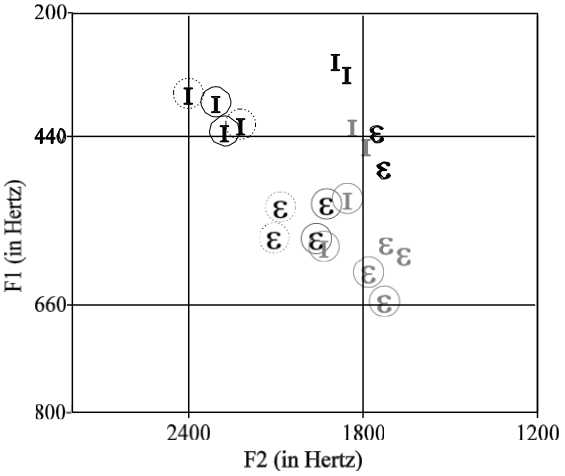

Figure 1: *Plot of Dutch vowels /ɪ/ and /ɛ/ used in the experiment. Vowels produced by a Northern Dutch female used in the training phase: Black circled. Novel Northern Dutch Female (speaker change): dotted circle. Northern Dutch male: Black without circle (sex change). Flemish*

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.

Familiarization Phase Pon/Deet: 20 trials		
Training Phase: Female Dutch 60 Trials		
Test Phase 120 trials (100%)	With Feedback: *Trained*: 72 trials (60%)	
	Without Feedback: 48 trials (40%):	
	24 Trained Stimuli (12 /ɪ/ and 12 /ɛ/)	**24 Novel Stimuli (12 /ɪ/ and 12 /ɛ/)**

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).

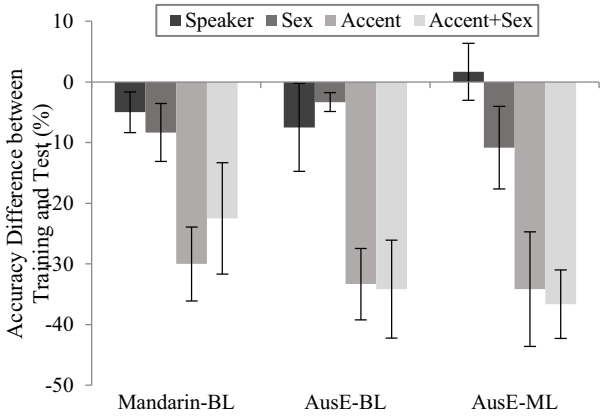

Figure 2. *Accuracy difference scores: negative scores indicate a decrease in accuracy at the test phase (compared to the training phase).*

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.