Acoustic distance explains speaker versus accent normalization in infancy

Paola Escudero, Karen E. Mulak & Samra Alispahic

The MARCS Institute, University of Western Sydney

paola.escudero@uws.edu.au, k.mulak@uws.edu.au, s.alispahis@uws.edu.au

Abstract

Acoustic/phonetic differences exist in cross-speaker and cross-accent speech. Young infants generally recognize speech across speakers but not across speakers of different accents. We examined how Australian English infants discriminated Dutch vowels produced by two speakers of the same accent, and by two speakers of two different accents. Acoustic analysis showed that the acoustic distance between same-vowel tokens produced by speakers of different accents was larger than between those produced by speakers of the same accent. Infants demonstrated greater difference in looking time to an accent than a speaker change, indicating that they noticed a difference in a vowel produced in a different accent more than one produced by another speaker with the same accent. This supports the hypothesis that acoustic distance underlies the relative ease in handling speaker versus accent variation.

Index Terms: speaker variation, accent variation, speech normalization, language development, vowel perception

1. Introduction

An utterance produced by two different people will vary considerably on acoustic/phonetic dimensions, due to individual differences in vocal tract characteristics. This variation between pronunciations is further increased when two speakers speak different regional accents of a language, as regional accents typically contain systematic variation in pronunciations. From experience, we know that adults are readily able to understand speech across both the variability seen among speakers of the same regional accent and the variability observed in regionally accented speech. But are these two types of variation handled to the same extent?

Recent studies on infant populations have found that like adults, young infants aged 7.5-9 months are able to recognize words across pronunciations by speakers with similar [1] and dissimilar [2], [3] (but see [1]) voices. This ability also extends to recognition of syllables [4], [5] and vowels [6] across speakers. Notably, chinchillas, budgerigars and zebra finches have also been found to be able to recognize vowels and words across speakers [7]–[11], suggesting that this ability to normalize speech across speakers and ascertain the relevant linguistic information may be innate.

Unlike adults, young infants are unable to recognize speech produced across accents. For instance, American infants familiarized to words produced by a speaker of American English did not recognize those words in passages produced by a Spanish-accented speaker of English [2], or a speaker of Canadian English [12], or vice versa

Indeed, even adults' ability to recognize speech across accents is subject to exposure. Adults were able to recognize words in Spanish- and Chinese-accented English as quickly as in non-accented English after receiving a short exposure to the accent [13], and different amounts of exposure are needed for successful processing of an accent depending on the accent and context [14]. With respect to vowels, an artificial regional accent study has shown that adults classified "wetch" as "witch" after 20 minutes of exposure to dialog that included the vowel shift of /ɪ/ to /ɛ/ [15], but did not make this classification without exposure.

Thus, it seems that without previous exposure, young infants, adults, and non-human mammals are able to recognize the invariant linguistic information of speech sounds even when produced by different speakers. This suggests that this ability is pre-linguistic and innate. By contrast, the ability to recognize speech across accents is not present in young infants, and adults require pre-exposure to words in the new accent. This suggests that accent normalization is an emergent ability that occurs at the lexical level, as it requires lexical exposure.

This lexical tie to resolving accent variation is also evident in older infants: Nineteen-month-olds exposed to words containing a vowel shift (e.g., "dog" shifted to "dag") subsequently treated the shifted pronunciation as a valid label for an associated visual referent (i.e., accepted "dag" as a label for a picture of a dog; [16]). As well, 15-month-olds' ability to recognize words across accents has also been shown to positively correlate with their expressive vocabulary score, further demonstrating the lexical link [17].

But why is it that resolving accent variation and resolving speaker variation appear to be handled in different ways? It may be that our auditory system (and that of vertebrates) is able to do away with the absolute physical differences between the productions of different speakers. In that respect, it has been shown that while there is considerable variation between the absolute F1 and F2 values of an individual vowel when produced by different speakers, this variation disappears when F1 and F2 are divided by higher formants such as F3 [18]. It may be that humans and vertebrates pay attention to these ratios rather than to absolute formant values when discriminating speech sounds, enabling rapid speaker normalization. Magneto-encephalographic studies have shown sensitivity in the auditory cortex to F1/F3 at latencies consistent with low level processing (between feature extraction and abstract processing, [19]). Thus there exists a plausible mechanism for rapid or even automatic speaker normalization within an accent.

Alternatively, it may simply be the case that the acoustic variation across the productions of different speakers from the same accent is smaller than that of speakers of different accents. In that respect, many cross-accent studies have shown that variability in the production of the same vowels across accents influences cross-accent, non-native and second-language vowel perception (for a recent review see the introduction and discussion in [20]).

As no previous study has directly compared how young listeners handle speaker versus accent variation in the production of vowels, the present study aimed to ascertain whether speaker variation is indeed less noticeable than accent variation in infancy. Given that the relative ease of speaker versus accent variation may change as a function of listeners' lexicon size, we compared responses across younger infants (<9 months) who have a very minimal receptive lexicon (<50 words) and an expressive lexicon of only 0-2 words, with older infants and toddlers (>11 months), who have a larger receptive vocabulary (77 to >250 words) and are fast developing their expressive vocabulary (8 to >87 words; [21]).

We examined how Australian English infants and toddlers recognize pronunciations of the vowel /ɪ/ spoken in a non-native language (Dutch) and across a different speaker of the same accent and across a different speaker of a different accent. As familiarity may play a role in speech normalization, we used non-native speech, so that both the voices and accents are unfamiliar to young listeners. If acoustic distance is responsible for the relative ease in normalizing speaker variation, we predicted that both age groups would be more likely to notice that a vowel has been produced by another speaker with a different accent than by another speaker with the same accent.

2. Method

2.1. Participants

Participants were 33 infants from Sydney, Australia. Participants were divided into two age groups: the "younger" group comprised 13 infants (8 females) aged 6.6 to 8.7 months (*M* = 7.8, *SD* = 0.6). The "older" group comprised 20 infants (11 females) aged 11.3 to 20.6 months (*M* = 15.4, *SD* = 3.7). Data from a further 22 infants were collected but not included due to fussiness (*N* = 18) or technical issues (*N* = 4).

2.2. Stimuli

Stimuli were naturally produced vowel tokens of the vowels /ɪ/ (as in KIT) and /ɛ/ (as in DRESS), excised from running speech produced by two female native speakers of Standard Northern Dutch spoken in North Holland (ND1, ND2) and one female native speaker of Standard Southern Dutch spoken in Flanders (FD).

Eight familiarization trials and four types of test trials (same, vowel change, speaker change, and accent change) were presented. Each trial was 13 s long and had 16 vowel tokens, with 750 ms between tokens. For the eight familiarization trials, two tokens of the vowel /ɪ/ produced by one speaker of North Holland Dutch alternated in a random order that was unique to each familiarization trial for a total of eight repetitions each (Table 1).

Table 1. *Vowel tokens for the familiarization and test trials. Tokens were produced by one of two female speakers of Standard Dutch (ND1, ND2) or a female speaker of Flanders Dutch (FD).*

Trial	Token alternation
Familiarization	/ɪ/-ND1, /ɪ/-ND1
Same	/ɪ/-ND1, /ɪ/-ND1
Vowel change	/ɪ/-ND1, /ɛ/-ND1
Speaker change	/ɪ/-ND1, /ɪ/-ND2
Accent change	/ɪ/-FD, /ɪ/-FD

The same trial was identical to a familiarization trial, while the three change trials had one token of a familiarization trial alternating with a token of the corresponding change, as shown in Table 1. For all trials, the two /ɪ/ tokens were arranged in a fixed random order. Vowel change trials had the vowel /ɛ/ produced by the same speaker as in familiarization as the alternating token, while speaker and accent change trials had tokens of /ɪ/ produced by a second female speaker of North Holland Dutch (ND2) and by a female speaker of Flemish Dutch (FD) respectively.

Absolute formant values (F1 and F2 in Hertz) for the tokens described in Table 1 can be seen in Figure 1. As mentioned in the introduction, a possible mechanism for speaker normalization may be via attending to F1/F3 and F2/F3 vowel formant ratios. The ratios for the vowel tokens used in these stimuli can be seen in Figure 2.

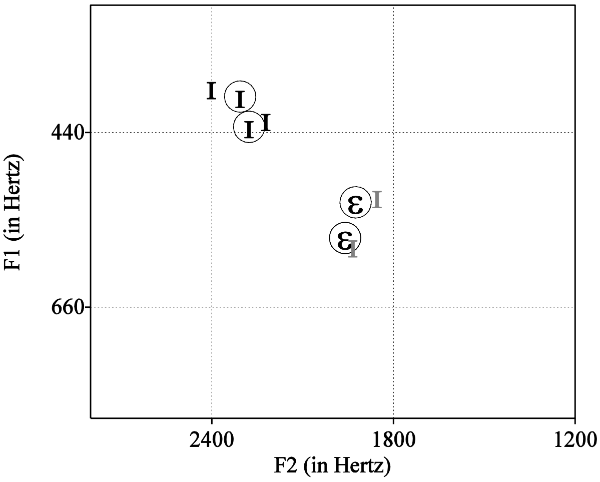

Figure 1. *F1 and F2 values of the stimuli used in the present study: North Holland speakers (ND, black), Flanders speaker (FD, grey). Vowel tokens from the speaker used in familiarization (ND1) are circled.*

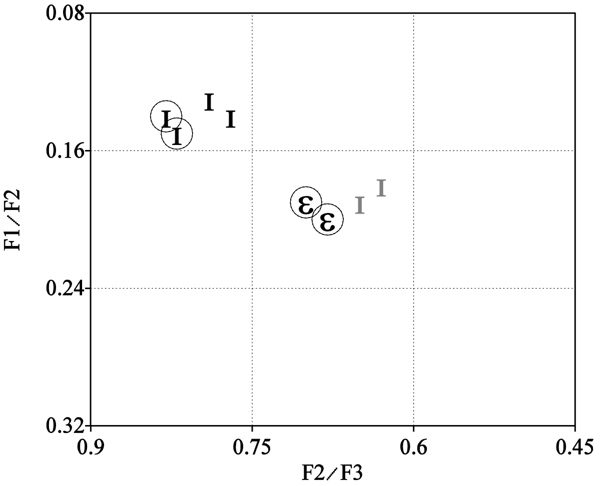

Figure 2. *F1/F3 and F2/F3 formant ratios for the stimuli in Figure 1. ND speakers (black), FD speaker (grey). Vowel tokens from the speaker used in familiarization (ND1) are circled.*

As can be observed through comparing the two figures with absolute formant values and ratios, Dutch /ɪ/ produced with a Flemish accent is as different from North Holland /ɪ/ than is North Holland /ɛ/. If infants pay attention to acoustic differences and are unable to normalize for accent distinctions such as the one shown in Figures 1 and 2, they should notice a change in /ɪ/ tokens from ND1 to FD more than a change from ND1 to ND2. Additionally, according to the acoustic distances in the two figures, the change from ND1 /ɪ/ to FD /ɪ/ (accent change) should be as noticeable as a change from ND1 /ɪ/ to ND1 /ɛ/ (vowel change).

2.3. Apparatus and Setup

Participants' gaze was recorded using a Tobii X120 eye-tracker (Tobii Technology, Danderyd, Sweden). Visual stimuli were presented on a 19-in Phillips LCD monitor. The eye-tracker was positioned 20 cm below the monitor, with its back edge 40 cm in front of the monitor. Auditory stimuli were played through an Edirol MA-15D speaker which was placed 57 cm to the right of the monitor. A QuickCam Orbit AF camera was placed on top of the monitor, allowing the experimenter to view participants from the adjoining room and verify that participants' gaze was being tracked when they were oriented toward the screen. Stimuli were presented using the E-Prime Extension for Tobii (E-Prime version 2.0.8.22; Psychology Software Tools, Sharpsburg, PA, United States).

2.3.1. Procedure

Participants were seated on their caretaker's lap and positioned so that their eyes were 70 cm from the front of the eye-tracker. For the duration of the study, caregivers wore MTUNE headphones (Macally, Ontario, Canada) that played a mixture of music and speech, and were instructed to look down or to the side during the experiment.

Following eye-tracking calibration, each participant completed a serial preference task that consisted of a familiarization phase and subsequent test phase. Each trial began with presentation of a dynamic cartoon paired with a non-linguistic sound that was positioned in the center of the monitor. This attention getter played until participants' gaze was oriented to the monitor.

Each participant completed eight familiarization trials and three test trials. Participants in Conditions 1 and 2 were presented with a non-alternating and vowel change test trial. Condition 1 also included a speaker change test trial, while Condition 2 also included an accent change test trial. For each familiarization and test trial, a colorful target appeared on the screen while the auditory stimuli played. Children's looking time to the colorful target during each trial was measured.

3. Results

The average of participants' looking time to the screen for the last two familiarization trials was subtracted from their total looking time to the screen for each test trial (Figure 3). These difference scores were analyzed in a 3 (test trial: same, vowel, speaker/accent change trials combined) x 2 (age group: younger infants, older infants) x 2 (condition: 1 [speaker change], 2 [accent change]) ANOVA with test trial as the within-subject variable. This analysis revealed a main effect of test trial, $F(2, 58) = 3.17$, $p = .049$, $\eta_p^2 = 0.1$. Participants had a larger difference in looking time for the combined speaker/accent change trials than for same trials, $F(1, 29) = 7.05$, $p = .013$, $\eta_p^2 = 0.20$. There was also a main effect of age group, $F(1, 29) = 5.90$, $p = .022$, $\eta_p^2 = 0.17$. Participants in the older age group had overall larger difference scores than younger infants.

To test our specific hypothesis that participants would show a greater difference in looking time for accent than for speaker change trials, we compared difference scores to these trials in an independent samples *t*-test. As can be seen in Figure 3, participants had a greater difference in looking time relative to the average of the last two familiarization trials for accent than for speaker change trials, $t(31) = 2.05$, $p = .049$, 95% CI [0.01, 3.27]. There was no difference across age group. Further independent samples *t*-tests revealed no differences between difference in looking time for same or vowel trials, and no differences across age groups.

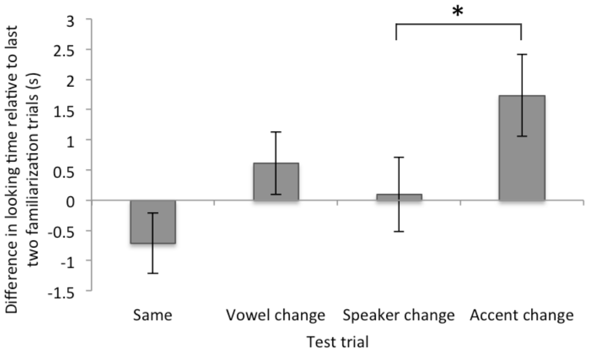

Figure 3. *Difference in looking time to test trials relative to the last two familiarization trials. Same and vowel change trials collapsed across conditions. Error bars represent one standard error.*

4. Discussion

The main question underlying this study was whether normalizing vowels across different speakers is easier than normalizing vowels across differing regional accents. As a first step in answering this question across infants and toddlers, we asked whether a smaller acoustic distance was responsible for the ease in normalizing speaker variation. Specifically, we tested whether the larger acoustic distance between tokens of the same vowel spoken by speakers of different accents would make an accent change more noticeable than speaker change. Our results show that indeed infants and toddlers were more likely to notice that a vowel was produced by a different speaker with a different accent than by a different speaker of the same accent.

Acoustic analysis of our stimuli showed that variation across tokens of the same vowel produced by speakers of the same accent was much smaller than that of tokens produced by speakers of different accents. Therefore, our results demonstrate that infants noticed the most salient acoustic difference, which indicates that speaker variation is easier to disregard because it corresponds to a smaller acoustic difference than variation in accent. This is the case because all speakers in our stimuli were females. It is yet to be shown whether comparable acoustic differences across speakers and accents would yield similar results. For instance, gender differences (female versus male vowel productions) have been shown to yield large differences that are comparable to accent variation, to the extent that vowels produced by males and females have very different absolute F1 and F2 values. It remains to be shown whether infants notice gender variation more than speaker variation and to the same extent as accent variation.

Although no significant age differences were found in the analysis, findings for the vowel change may suggest that only the data for the older group were reliable. That is, the older children had a higher increase in looking to vowel change trials (*M* = 1.39 s, *SD* = 3.00 s) than did younger infants (*M* = -0.58 s, *SD* = 2.62 s). This suggests that only the older infants' looking times are based on the acoustic differences between the vowel tokens in the test trials. In addition, inspection of the younger group's data shows that their difference scores for change test trials are close to zero. A larger number of participants in each age group will clarify this potential age difference. Additionally, comparisons with how native Dutch infants perform the same task would indicate whether experience with Dutch vowels and Dutch voices plays a role in the handling of speaker versus accent variation at this early age.

In sum, our results are a first step in explaining why very young listeners may be able to handle speaker variation more easily than accent variation. This is because infants and toddlers noticed that a vowel was produced by a speaker of a different accent more readily than when it was produced by a speaker of the same accent. Since all speakers that produced the stimuli were females, the explanation lies in the larger acoustic difference across tokens of speakers with different accents than across tokens of speakers with the same accent. Further research comparing gender and accent as well as native and non-native listeners will show whether acoustic distance or other mechanisms such as the computation of formant ratios underlies the ease with which infants normalize speaker variation.

5. References

[1] D. M. Houston and P. W. Jusczyk, "The role of talker-specific information in word segmentation by infants," *J. Exp. Psychol. Hum. Percept. Perform.*, vol. 26, no. 5, pp. 1570–1582, 2000.

[2] R. Schmale and A. Seidl, "Accommodating variability in voice and foreign accent: Flexibility of early word representations," *Dev. Sci.*, vol. 12, no. 4, pp. 583–601, 2009.

[3] M. Van Heugten and E. K. Johnson, "Infants exposed to fluent natural speech succeed at cross-gender word recognition," *J. Speech Lang. Hear. Res.*, 2011.

[4] G. Dehaene-Lambertz and M. Pena, "Electrophysiological evidence for automatic phonetic processing in neonates," *Neuroreport*, vol. 12, no. 14, pp. 3155–3158, 2001.

[5] P. W. Jusczyk, K. Hirsh-Pasek, D. G. Kemler Nelson, L. J. Kennedy, A. Woodward, and J. Piwoz, "Perception of acoustic correlates of major phrasal units by young infants," *Cognit. Psychol.*, vol. 24, no. 2, pp. 252–293, Apr. 1992.

[6] P. K. Kuhl, "Perception of auditory equivalence classes for speech in early infancy," *Infant Behav. Dev.*, vol. 6, no. 2–3, pp. 263–285, 1983.

[7] C. K. Burdick and J. D. Miller, "Speech perception by the chinchilla: discrimination of sustained ‖a‖ and ‖i‖," *J. Acoust. Soc. Am.*, vol. 58, no. 2, pp. 415–427, Aug. 1975.

[8] R. J. Dooling and S. D. Brown, "Speech perception by budgerigars (Melopsittacus undulatus): Spoken vowels," *Percept. Psychophys.*, vol. 47, no. 6, pp. 568–574, Nov. 1990.

[9] K. Kluender, R. Diehl, and P. Killeen, "Japanese quail can learn phonetic categories," *Science*, vol. 237, no. 4819, pp. 1195 –1197, 1987.

[10] P. K. Kuhl and J. D. Miller, "Speech perception by the chinchilla: voiced-voiceless distinction in alveolar plosive consonants," *Science*, vol. 190, no. 4209, pp. 69–72, Oct. 1975.

[11] V. R. Ohms, P. Escudero, K. Lammers, and C. Ten Cate, "Zebra finches and Dutch adults exhibit the same cue weighting bias in vowel perception," *Anim. Cogn.*, Jul. 2011.

[12] R. Schmale, A. Cristia, A. Seidl, and E. K. Johnson, "Developmental changes in infants' ability to cope with dialect variation in word recognition," *Infancy*, vol. 15, no. 6, pp. 1–13, 2010.

[13] C. M. Clarke and M. F. Garrett, "Rapid adaptation to foreign-accented English," *J. Acoust. Soc. Am.*, vol. 116, no. 6, p. 3647, 2004.

[14] C. Floccia, J. Goslin, F. Girard, and G. Konopczynski, "Does a regional accent perturb speech processing?," *J. Exp. Psychol. Hum. Percept. Perform.*, vol. 32, no. 5, pp. 1276–1293, Oct. 2006.

[15] J. Maye, R. N. Aslin, and M. K. Tanenhaus, "The weckud wetch of the wast: Lexical adaptation to a novel accent," *Cogn. Sci.*, vol. 32, no. 3, pp. 543–562, Apr. 2008.

[16] K. S. White and R. N. Aslin, "Adaptation to novel accents by toddlers," *Dev. Sci.*, vol. 14, pp. 372–384, 2010.

[17] K. E. Mulak, C. T. Best, M. D. Tyler, C. Kitamura, and J. R. Irwin, "Development of phonological constancy: 19-month-olds, but not 15-month-olds, identify words spoken in a non-native regional accent," *Child Dev.*, vol. 84, no. 6, pp. 2064–2078, 2013.

[18] L. Deng and D. O'Shaughnessy, *Speech Processing: A Dynamic and Optimization-Oriented Approach*. CRC Press, 2003.

[19] P. J. Monahan and W. J. Idsardi, "Auditory sensitivity to formant ratios: Toward an account of vowel normalisation," *Lang. Cogn. Process.*, vol. 25, no. 6, pp. 808–839, 2010.

[20] P. Escudero, B. Sisinni and M. Grimaldi, "The effect of vowel inventory and acoustic properties in Salento learners of Southern British English vowels", *Journal of the Acoustical Society of America* **135**, 1577-1584 (2014)

[21] P. S. Dale and L. Fenson, "Lexical development norms for young children," *Behav. Res. Methods Instrum. Comput.*, vol. 28, no. 1, pp. 125–127, 1996.