PERCEPTUAL SEPARATION OF SPECTRALLY OVERLAPPING VOWELS

Robert A. Fox and Ewa Jacewicz

Department of Speech and Hearing Science, The Ohio State University, Columbus, OH, USA
fox.2@osu.edu, jacewicz.1@osu.edu

ABSTRACT

Previous research on vowel identification reported excellent intelligibility of spectrally adjacent vowels despite reduced acoustic separation in their formant frequencies and extensive talker variability. The current study questioned this status quo by introducing several sources of real-world variation in vowel production due to variable stress patterns, sound change, and talker age and gender. Controlling for talker and listener dialect, the study was conducted in a small Appalachian community in Southern United States. As expected, identification rates were lower and more variable across individual vowel categories when compared with previous reports that used citation-form stimuli. Perceptual separation of mid and low vowels was more challenging for listeners than separation of high and mid vowels, which was also reflected in the increased number of confusions with the neighbors. The study provides new evidence that the identities of neighboring vowels can be compromised when additional sources of variation alter their spectro-temporal properties.

Keywords: Vowel perception, sociophonetics, dialect, talker variability, Appalachian English

1. INTRODUCTION

In one of the most influential studies on the acoustics and perception of vowels, Peterson and Barney discovered a surprising lack of correspondence between the acoustic overlap among neighboring vowels and their high perceptual separability [17]. Decades later, very similar results were obtained in a modern replication of this study by Hillenbrand et al. [6], who again reported high identification of adjacent vowels (such as /ɛ/ and /æ/) despite very poor acoustic separation in their formant frequencies (F1 and F2). Importantly, the excellent intelligibility reported in both studies did not seem to be hampered by extensive variability in pronunciation patterns from one talker to the next, as stimulus material was produced by 76 talkers (men, women, and children) in [17] and 139 talkers in [6].

Over the years, many possibilities have been explored with regard to which cues can aid listeners' identification of adjacent vowels under extensive talker variability (review in [7]). Presumably, listeners benefit most from a combination of cues including vowel duration, fundamental frequency, and dynamic formant pattern termed vowel-inherent spectral change [16]. However, it is still unknown how these and other secondary cues [13] interact to inform listeners' decisions when the pronunciation of vowels is additionally altered by real-world variations such as sentence prosody, speaking rate, and a range of sociophonetic variables including regional variation and diachronic sound change. All these sources of variation are likely to obscure identification of spectrally overlapping vowels, particularly when lexical cues provide limited support.

The current study sought to establish the viability of perceptual separation of neighboring vowels under such extensive variations, focusing on the combined effects of variable stress, sound change, and talker age and gender. The study controlled for dialect so that both talkers and listeners came from the same speech community and spoke a local variety of Appalachian American English. We predicted that identification accuracy in our study would not be as high as in [17, 6] because of the differences in the nature of phonetic variation in the stimulus material. While evidence for the excellent separability of vowels in [17, 6] comes from citation-form utterances, listeners in the current study were presented with a competing set of demands imposed by extensive variation in vowel characteristics. We also expected more variability in identification rates across individual vowel categories as some of the vowels in our study were involved in an ongoing sound change.

2. METHODS

Five vowels were of interest: /ɪ, e, ɛ, æ, ai/. In the Appalachian variety studied here, these vowels are not only overlapping because of their extensive formant movement known as Southern Breaking (in /ɪ, ɛ, æ/), but they also participate in the Southern Shift, a chain-like rotation of front vowels involving the monophthongization of /ai/, the centralization of /e/, the peripheralization of /ɪ, ɛ/ [14, 3], and raising of /æ/ in older speakers [9]. In addition, the nature of the spectral overlap found in older generations is changing in children due to increased influence of

mainstream American English [11]. All these patterns are shown in Figures 1-4.

2.1. Stimuli

The tokens *bids, bades, beds, bads*, and *bides* containing the vowels /ɪ, e, ɛ, æ, ai/, respectively, were produced by 40 talkers, 20 adults aged 50-65 years (10 males, 10 females) and 20 children aged 9-12 (10 boys, 10 girls). All talkers spoke a local Appalachian English variety typical of the dialect region Inland South in western North Carolina. The tokens were produced in experimentally constructed sentences with variable stress patterns. Each talker contributed 10 unique exemplars of all five words that either carried the main sentence stress (5 tokens) or were unstressed (5 tokens). The final stimulus set for perceptual testing consisted of 400 unique tokens excised from read sentences. Average dynamic formant patterns of adults' and children's vowels measured in these words are displayed in Figures 1-4. To allow comparisons, formant values were normalized using the Lobanov's procedure [15] on the basis of a 14-vowel set produced by each talker.

Figure 1: Average formant patterns sampled at 5 time-points (20-35-50-65-80%) in male adults

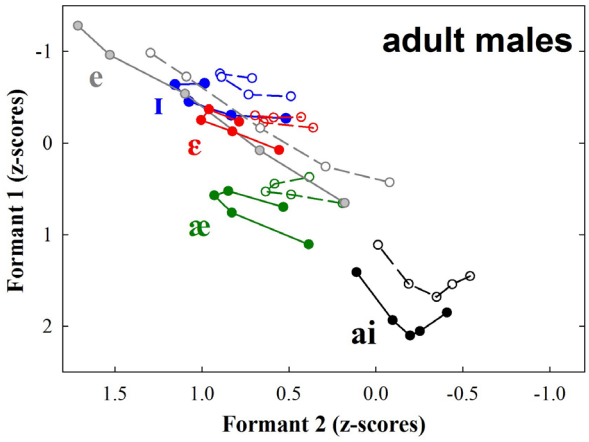


Figure 2: Average formant patterns in female adults

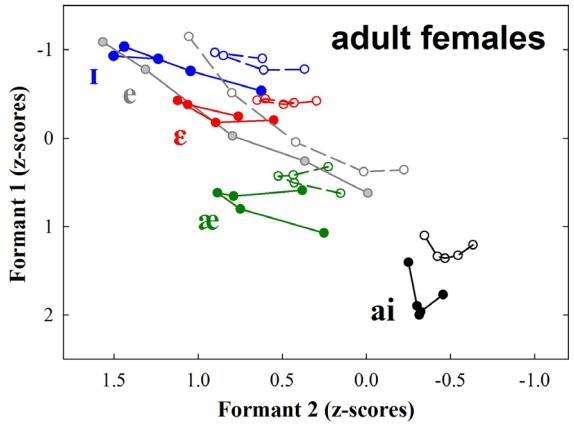


Figure 3: Average formant patterns in boys

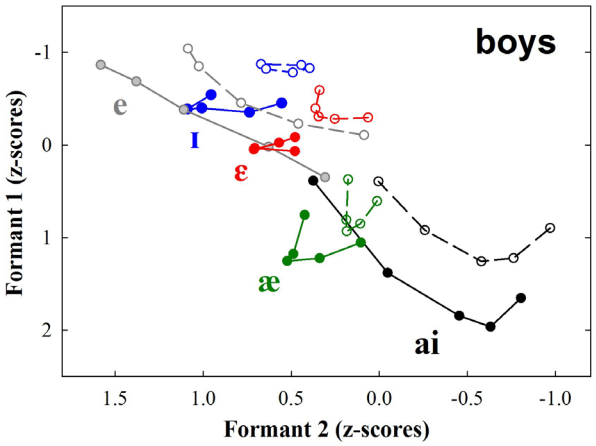
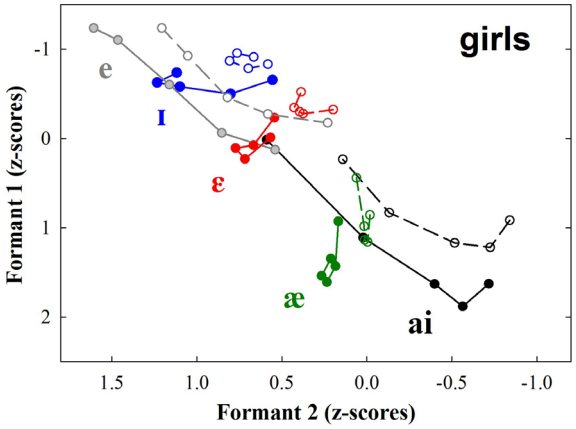


Figure 4: Average formant patterns in girls



2.2. Listeners and procedure

Listeners were 16 middle-age adults, ranging from 43-59 years (M = 52.75, SD = 5.46), 14 females and 2 males. All listeners were born and raised in the local community and have never left the area other than for occasional trips. All were employed and pursued a variety of professional careers, and have never participated in research experiments before. None reported hearing loss or any hearing problems.

The experiment was conducted in a quiet room at Western Carolina University in Cullowhee, NC. Each listener was tested individually. Signals were delivered over Sennheiser HD600 headphones at a comfortable listening level. Prior to presentation, all tokens were equalized for mean intensity.

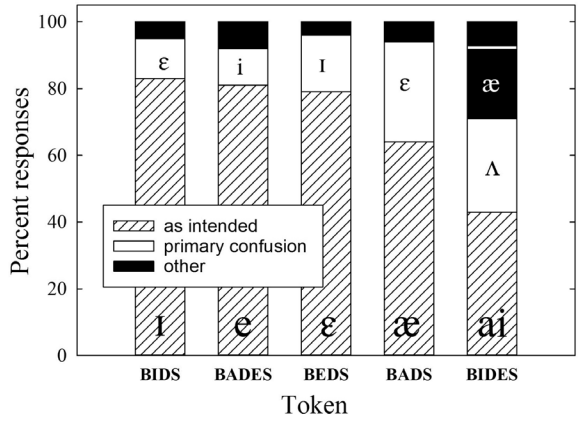
The stimuli were presented in random order in two blocks of 200 tokens each. A 20-item practice first familiarized the listeners with the task and ensured that they were able to match the orthographic form with the sound. The tokens in the practice trial were different than those in the experiment. The listeners were told that after listening to each word, they were to decide which word was played by selecting one of

the seven boxes on the computer monitor: beads, bids, bades, beds, bads, bides, and buds. Only one repetition was allowed and the listeners were asked to guess if still uncertain which response to choose. The experiment was self-paced. A custom program written in MATLAB controlled data collection.

3. RESULTS

Overall identification rates (IDRs, in % correct) by vowel category along with predominant confusions are shown in Figure 5. Accuracy was highest for the high vowel /ɪ/ and gradually declined for the mid /e, ɛ/ and the low vowels /æ, ai/, respectively. The observed decline for the mid and low vowels is in line with [6]. We did not compare our results with [17] because the vowel /e/ was not included in their set. As predicted, the IDRs in the current study were lower than in [6]. The differences for /ɪ, e, ɛ, æ/ were 16, 17, 16, and 30%, respectively. The IDR for /ai/ could not be compared with either [17] or [6] because the vowel was not included in their sets. Here, it was the monophthongal production of /ai/ that resulted in a particularly low IDR (43%) and extensive confusions. This outcome was not unexpected because low accuracy for /ai/ (53%) was previously reported in another study of this dialect [10].

Figure 5: Overall accuracy and confusion of vowels

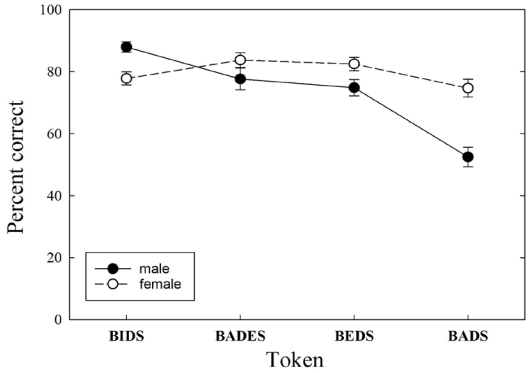


The confusion pattern seems straightforward. Based on their acoustic proximity, the vowels /ɪ, ɛ/ were confused with one another, and /æ/ was confused with /ɛ/. The confusions of /e/ with /i/ are most likely due to their similar patterns of spectral dynamics as the vowel /i/ is diphthongized in older speakers of this dialect and the direction of formant movement corresponds to that in /e/ [3]. Finally, the monophthongal /ai/ was primarily confused with the monophthong /ʌ/. The second dominant confusion was with /æ/, as indicated in Figure 5.

Listeners' correct identification responses were analysed using linear mixed-effects models in IBM SPSS Statistics (version 25, 2017) [8]. Following recent recommendations for analysing proportional data in a forced-choice task [12, 18], no arcsine transform was applied to analyze the response proportions. The best-fitting model was chosen using forward selection, adding one predictor at a time starting with a baseline model that only included the intercept. Vowel, stress, talker age group, talker gender and their interactions were entered as fixed effects. Listener was a random effect. We used log-likelihood comparisons to determine the significance of the fixed effects.

The initial model was overly complex and included three-way interactions due to a differential response pattern for /ai/. To keep the model parsimonious, /ai/ was excluded and responses for *bides* were analysed separately. A much simpler model for /ɪ, e, ɛ, æ/ revealed a significant main effect of vowel (χ^2 (3)=70.54, p<.001). Subsequent pairwise comparisons showed that accuracy for /æ/ was significantly lower when compared with any other vowel (p<.001) and the differences between the remaining three vowels were not significant. The main effect of stress was significant (χ^2 (1)=45.86, p< .001); accuracy was higher for stressed vowels than for unstressed (82 vs 71%). The main effect of gender was also significant (χ^2 (1)=15.27, p<.001) with higher accuracy for female speakers than for male speakers (80 vs. 73%). The main effect of age group did not significantly improve the model (χ^2 (1)=0.05, p=.830) and was removed. The model was further improved by a significant vowel by gender interaction (χ^2 (3)=51.57, p<.001). As shown in Figure 6, its locus was in the significant gender-related variations for the vowels /ɪ/ (*bids*) and /æ/ (*bads*) as revealed by Tukey-HSD comparisons (p<.001 for both).

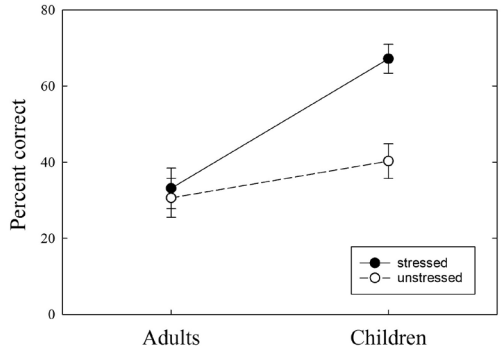
Figure 6: Accuracy by vowel and gender



The second separate model for /ai/ was constructed with the same fixed and random effects except that vowel was not included as a predictor. The best-fitting model revealed a significant main effect of age group (χ^2 (1)=40.83, p<.001); the accuracy was higher for children's vowels than for adult vowels. Two interactions with age group were also included.

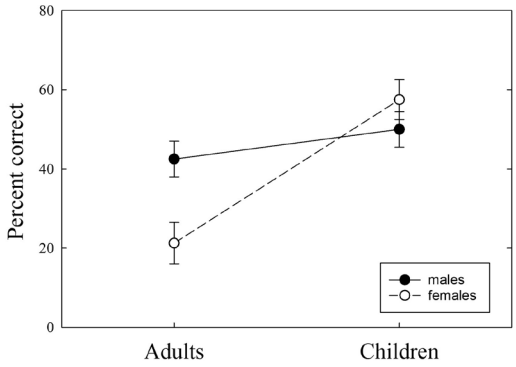
The significant age group by stress interaction (χ^2 (1)=16.44, p<.001), shown in Figure 7, arose because stress did not affect IDRs for adults (p=.476) but it did for children, with higher accuracy when the vowel was stressed (p <.001).

Figure 7: Age group by stress interaction for *bides*



The second significant interaction was between age group and gender (χ^2 (1)=29.52, p <.001). As shown in Figure 8, accuracy for adult male vowels was higher than for female vowels (p<.001) but the opposite was true for children's vowels (p<.001). Overall, the second model revealed that the differential pattern for /ai/ was mostly due to the significant effects of age group.

Figure 8: Age group by gender interaction for *bides*



4. DISCUSSION

The current study found that identities of neighboring vowels can be compromised when additional sources of variation alter their spectro-temporal properties. The sources examined here included linguistic stress, sound change in a speech community (in apparent time, comparing older adults and children), and talker gender. For the front vowels /ɪ, e, ɛ, æ/, accuracy for stressed vowels was higher than for unstressed, which is consistent with previous research [5, 1, 2]. Accuracy was also significantly higher in response to female rather than male talkers, supporting the perceptual advantage of female speech [6, 10, 4]; however, gender-related differences varied across individual vowels. The lack of a significant effect of age group or interactions with age group is somewhat surprising given the generational differences in spectral overlap of the front vowels and other changes such as the lowering and reduction of formant movement in /æ/ in girls. On the other hand, the local listeners were likely familiar with variable pronunciation patterns in this speech community and could thus make perceptual adjustments when responding to old and new pronunciation forms.

However, the differential pattern for /ai/ demonstrated that generational variations can have a profound effect on accuracy. The confusions revealed that, when hearing the monophthongal /ai/ (whether stressed or unstressed) listeners found the distinctions between *bides, buds,* and *bads* challenging, and only the increased formant movement in children (more so in girls and more so when the vowel was stressed) was able to disambiguate the signal and aid in making perceptual decisions. This example provides evidence that spectral dynamics can play a critical role in vowel identification when other cues become unreliable.

Importantly, the relationship between accuracy and vowel height was more transparent in the current study than in [6], namely that accuracy gradually declines with each vowel descending in height. Intrigued by this trend, we found a correspondence between the decline in accuracy and the frequency of the second stimulus repetition option that was available to the listeners. In particular, out of a total of 3.01% of all stimulus tokens that were heard twice, the distribution of repetitions was: 0.33% (*bids*), 0.48% (*bades*), 0.59% (*beds*), 0.70% (*bads*), and 0.91% (*bides*). The increase in listener uncertainty with each descending vowel category suggests that, irrespective of the amount of spectral overlap, perceptual separation of mid and low vowels is more challenging than separation of high and mid vowels, which is also reflected in the increased number of confusions with the neighbors.

The current study also suggests that greater spectral overlap of neighboring vowels does not necessarily lead to their increased confusions as demonstrated by the results for /ɪ/ and /ɛ/. Although extensive acoustic variations did obscure their identification rates when compared with citation-form speech [17, 6], listeners were quite successful in separating the vowels. Perceptual strategies that led to this success are still unknown and need to be uncovered in the future.

5. ACKNOWLEDGMENTS

This work was supported by NIH/NIDCD Grant R01DC006871. Special thanks to Janaye Houghton for help with data collection.

6. REFERENCES

[1] Cutler, A. 1986. Forbear is a homophone: Lexical prosody does not constrain lexical stress. *Lang. Speech* 29, 201-220.

[2] Cutler, A. 2005. Lexical stress. In: Pisoni, D., Remez, R. (eds). *The Handbook of Speech Perception*. Hoboken, NJ: Wiley-Blackwell, 264-289.

[3] Farrigton, C., Kendall, T., Fridland, V. 2018. Vowel dynamics in the Southern Vowel Shift. *American Speech* 93, 186-222.

[4] Ferguson, S. H. 2004. Talker differences in clear and conversational speech: Vowel intelligibility for normal-hearing listeners. *J. Acoust. Soc. Am*. 116, 2365-2373.

[5] Fry, D B. 1955. Duration and intensity as physical correlates of linguistic stress. *J. Acoust. Soc. Am*. 27, 765-768.

[6] Hillenbrand, J., Getty, L. A., Clark, M. J., Wheeler, K. 1995. Acoustic characteristics of American English vowels. *J. Acoust. Soc. Am*. 97, 3099-3111.

[7] Hillenbrand, J. 2013. Static and dynamic approaches to vowel perception. In: Morrison, G. S., Assmann, P. F. (eds). *Vowel Inherent Spectral Change*. Berlin: Springer, 9-30.

[8] IBM SPSS Statistics, v. 25. 2017. International Business Machines Corp. Armonk, NY.

[9] Jacewicz, E., Fox, R., A., Salmons, J. 2011. Cross-generational vowel change in American English. *Language Variation and Change* 23, 45-86.

[10] Jacewicz, E., Fox, R. A. 2012. The effects of cross-generational and cross-dialectal variation on vowel identification and classification. *J. Acoust. Soc. Am*. 131, 1413-1433.

[11] Jacewicz, E., Fox, R. A. 2019. The old, the new, and the in-between: Preadolescents' use of stylistic variation in speech in projecting their own identity in a culturally changing environment. *Developmental Science* 2019;22:e12722, DOI: 10.1111/desc.12722

[12] Jaeger, T. F. 2008. Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *J. Mem. Lang.* 59, 434-446.

[13] Klatt, D. H. 1976. Linguistic uses of segmental duration in English: Acoustic and perceptual evidence. *J. Acoust. Soc. Am*. 59, 1208-1221.

[14] Labov, W., Ash, S., Boberg, C. 2006. *Atlas of North American English: Phonetics, Phonology, and Sound Change.* New York: Mouton de Gruyter.

[15] Lobanov, B. 1971. Classification of Russian vowels spoken by different speakers. *J. Acoust. Soc. Am*. 49, 606-608.

[16] Nearey, T. M., Assmann, P. F. 1986. Modeling the role of vowel inherent spectral change in vowel identification. *J. Acoust. Soc. Am*. 80, 1297-1308.

[17] Peterson, G. E., Barney, H. L. 1952. Control methods used in a study of the vowels. *J. Acoust. Soc. Am*. 24, 175–184.

[18] Warton, D I., Hui, F.K. 2011. The arcsine is asinine: The analysis of proportion in ecology. *Ecology* 92, 3-10.