SPEAKER-ADAPTATION TO /ɪ/ - /ɛ/ MERGER: AN EYE-TRACKING STUDY

Kiwako Ito & Kathryn Campbell-Kibler

Ohio State University, USA
ito@ling.ohio-state.edu; kbck@ling.ohio-state.edu

ABSTRACT

This study uses eye-tracking to investigate how a brief exposure to the /ɪ/ - /ɛ/ merger (between, e.g., *pin* and *pen*) affects subsequent lexical access. Merged and non-merged speakers gave instructions (e.g., "Click on the *pencil*") for an object search task. Each voice was paired with a photo of a White or Black face in professional or non-professional dress, counter-balanced across participants. The results showed a clear adaptation to merged speakers. A minor competition (between, e.g., *pencil* and *pins*) was observed regardless of voice for unambiguous /ɛn/-words in Block 1. After hearing tokens of /ɪn/-words (/ɛn/-like for merged speakers) in Block 2, listeners responded more slowly in Block 3 to /ɛn/-words from the merged speakers only. The effect of photo race suggests that listeners had trouble integrating non-merged voices with Black faces and merged voices with White faces. No effect of dress was found.

Keywords: vowel merger, speaker-adaptation, eye-tracking, speech processing, sociophonetics

1. INTRODUCTION

It is well established that listeners are capable of quickly accommodating to speaker-specific variation in pronunciation. Experimental studies have suggested that speaker-specific phonetic details, as well as multi-dimensional speaker representations (including social information like gender and age) are stored in listeners' memory and that these detailed perceptual traces contribute to subsequent speech processing [5, 7, 14]. A common source of cross-speaker variation is dialect variation, be it regional, racial or ethnic, or class-related, among other factors. While many studies on dialect-related speech perception have focused on how a lack of particular phonemic contrast in the native dialect may result in failure to discriminate sounds distinguished in another dialect [3, 6, 10], fewer studies have investigated how listeners adapt to speaker-specific dialectal cues in speech processing. Dahan, et al. [4] demonstrated that listeners quickly learned dialect-specific vowel raising in words like *bag* (where the /æ/ is raised to /ɛ/), and their detection of target words like *back* (with un-raised /æ/) was facilitated as a result of this rapid speaker-adaptation. The present study modifies the experimental paradigm of Dahan, et al. [4] and tests whether a brief experience with a dialect-specific vowel merger leads to a lexical competition as the result of speaker-adaptation. In addition, the experiment examines whether and how the adaptation process is influenced by social information, in the form of visual cues to race and socio-economic background.

The /ɪ/ - /ɛ/ merger, often referred to as the pin/pen merger, occurs when speakers merge these two vowels in pre-nasal environments. It is a well-studied feature originating in the Southern US, and is documented in letters and other writings as far back as the 18th century [12]. The feature is now well established throughout the South [9], although appears to be disappearing among younger White speakers in some urban centers [8]. African American speakers throughout the country show high amounts of the merger [9, 15], due to the relative recency of the Great Migration of African Americans from the South. Although the merger is reported to have gained prestige in Southern areas [2], Southern speech remains stigmatized and seen as incorrect in other parts of the country [13]. As a result, Ohio listeners are likely to associate the merger with African American speakers, and with less educated White speakers from the South.

In the present study, the instructions for object search are presented with a face photo of a Black or White male in either professional or casual dress, to test whether and how listeners integrate these sociolinguistic cues while learning speaker-specific pronunciations.

2. EXPERIMENT

2.1. Participants

68 undergraduate students at the Ohio State University participated for partial fulfillment of

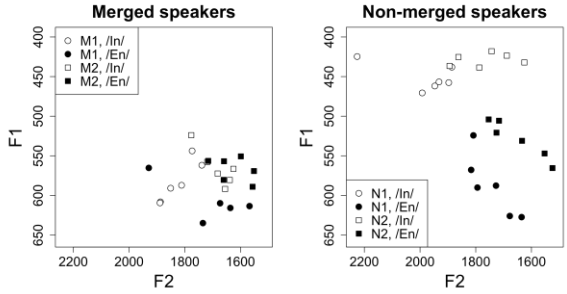
course credit. Data from 12 participants were excluded from the analysis due to system failure (4), and to being non-native speakers of American English (8). Among the 56 participants whose data were analyzed, 30 were from central and north Ohio, 6 were from southern Ohio, and the remaining 20 spent more than 5 years of their childhoods outside Ohio.

2.2. Materials

2.2.1. Auditory stimuli

Thirteen speakers were recorded at 44.1KHz using Praat, producing the instructions "*Click on the XXX.*" by naming labeled photos of objects one by one. Two merged (M1, M2) and two non-merged (N1, N2) voices were selected for their overall clarity and pronunciation patterns of the target word pairs. Fig. 1 shows each speaker's formant distributions of the target word pairs. In contrast to previous descriptions [1, 14], the two merged speakers showed lowering of /ɪ/ rather than raising of /ɛ/, making their /ɪn/-words (e.g., *pins*) momentarily ambiguous with their /ɛn/-words (e.g., *pencil*).

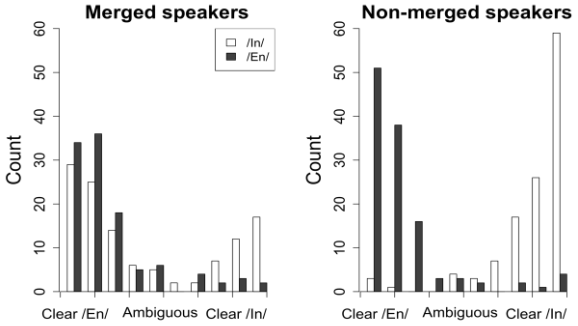
Figure 1: F1/F2 distributions of the target /ɪn/ and /ɛn/ words.



To select the auditory stimuli for the eye-tracking experiment, target word pairs were submitted to the visual analogue scale (VAS) rating task [11]. In this task, 10 participants (who did not participate in the eye-tracking experiment) heard the carrier phrase "*Click on the…*" plus the initial fragment of each word (e.g., /*p*ɛn/). They indicated whether they thought the fragment came from, e.g., *pencil* or *pins* and how certain they were by clicking on a line labeled at the ends with the two candidate words.

For the merged speakers the most /ɛn/-like token for each item was selected, yielding maximally ambiguous /ɪn/ tokens and minimally ambiguous /ɛn/ tokens. For non-merged speakers, the most distinct tokens were selected. Fig. 2 shows the VAS responses to the selected /ɪn/- /ɛn/ word pairs. While non-merged voices had clearly separated responses concentrating toward the two ends of the scale, the responses to merged voices showed more overlaps and confirmed that /ɪn/-words were often perceived as /ɛn/-words.

Figure 2: The results of VAS ratings on the target /ɪn/ and /ɛn/ words spoken by merged (M1, M2) and non-merged (N1, N2) speakers.



2.2.2. Visual stimuli

Each of the 60 critical slides contained one /ɪ/- /ɛ/ object pair (e.g., *pencil-pins*), a rhyme pair (e.g., *sneaker-speaker*), and four phonetically unrelated distractors (e.g., *bunk bed, sunglasses, drum set, swing*). An additional set of 96 slides was created for the filler trials that mentioned either one member of the rhyme pair (48) or one of the distractors (48) as the target object. The center cell had one of the four types of face photos (a White/Black male in professional/non-professional dress), yielding four versions of each slide.

Figure 3: An example display used in the eye-tracking study.

2.3. Design, procedure & predictions

Before the eye-tracking experiment, participants named each of the 48 photo objects twice, presented on the monitor one by one. This

familiarized participants with the names of the objects for the following object search task.

Participants' eyes were then calibrated using Clearview 5-point calibration function. Participants were randomly assigned to one of the four lists that combined each voice with one of four possible face photos. A total of 144 trials were presented in three blocks. In Block 1, three /ɛn/-words (*pencil, men, dentist sign*) were presented in each of the four voices. Having no previous experience with these speakers who all pronounce relatively clear /ɛn/, listeners' eye movements should show fast detection of the target objects with little fixations to the competitors regardless of voice. Block 2 presented six /ɪn/-words (*pins, bin, fins, mint, dinner plate, tin can phone*). The target detection was predicted to be delayed with merged voices that contained /ɛ/-like vowels. Block 3 presented new /ɛn/-words (*bench, fence, tent stake*) as well as the three /ɛn/-words already presented in Block 1. If listeners learned speaker-specific phonetic variation, their object detections for /ɛn/-words in Block 3 should be slower than those in Block 1 for the merged voices. This is because the listeners may momentarily consider the possibility that the /ɛ/-like vowel in the merged voices may in fact be part of the /ɪn/-words. For the non-merged voices, such delay in the target detection should not be observed if the listeners have learned that the vowels /ɪ/ and /ɛ/ are clearly separated in these speakers.

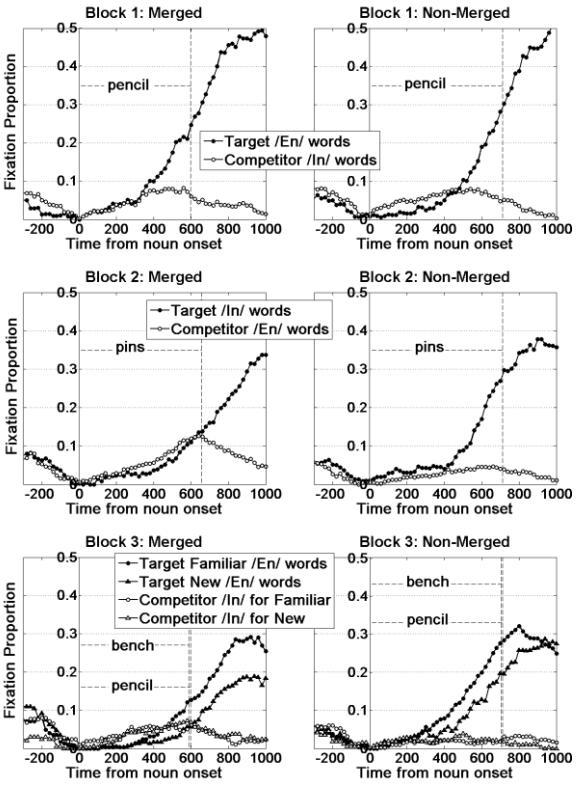
3. RESULTS

The fixation data were collected at 60Hz. For the statistical analysis, they were aggregated into 60ms time bins, and were submitted to mixed-effects logistic regressions which modeled the fixation likelihood with predicting factors such as time, voice, race, and dress and with random factors subject and item.

3.1. Effect of speaker-adaptation on lexical access

Participants' eye-movements showed a clear effect of their experience with the merger. Fig. 4 show the fixation proportions to the target and to the competitor for the merged and non-merged voice trials for each block (due to space limitations, these figures collapse four lists). As predicted, the clear pronunciation of /ɛn/-words in Block 1 did not lead to frequent fixations to the /ɪ/-word competitors in both merged and non-merged voices. The timing of target detection was somewhat delayed for the non-merged voices (voice*time coeff: .106, wald Z= 2, p<.05).

Figure 4: Fixation proportions to the target and competitor for Merged and Non-Merged voices.



In Block 2, a competition between the target /ɪn/-words and the competitor /ɛn/-words was higher for the merged than for the non-merged voices (compare empty lines in Fig. 4: middle panels. voice*time coeff: -.318, wald Z=-3.12, p<.01).

Most importantly, the experience with speaker-specific merger clearly modulated the responses to /ɛn/-words in Block 3. With the merged voices, listeners' detection of the repeated /ɛn/-targets was delayed compared to Block 1 despite their familiarity (block*time coeff: .029, wald Z=2.05, p<.05). The detection of the new /ɛn/-targets was even further delayed (block*time coeff: .057, wald Z=1.8, p=.069). With the non-merged voices, fixations to the repeated /ɛn/-targets were faster than in Block 1, probably due to their familiarity (block*time coeff: -.05, wald Z=-4.01, p<.0001). The fixations to new /ɛn/-targets were delayed as compared to the /ɛn/-targets in Block 1 (block*time coeff: -.08, wald Z= -3.06, p<.01).

3.2. Effect of race and outfit of speaker

The results also showed that social cues had a large impact on listeners' speech processing. In Block 1, being paired with a Black face delayed target detection, but only for non-merged voices (race*bin coeff: .11, wald Z=2.93, $p<.01$), not for the merged voices (race*bin coeff: .005, wald Z=1.5, $p=.13$). This suggests that listeners had difficulty integrating the non-merged voices with Black faces but had little trouble processing merged voices with White faces.

Interestingly, while listeners were experiencing the merger in Block 2, their fixations to the competitors were higher for the White than for the Black faces, regardless of the dress (voice*race*bin coeff: .36, wald Z=2.73, $p<.01$).

After the listeners experienced the merger, particular effects of race and outfit were not observed.

4. DISCUSSION AND CONCLUSION

The present results demonstrate that listeners rapidly adapt to speaker-specific variation in pronunciation and use their knowledge about the voices in the subsequent lexical processing. As predicted, the merger introduced an ambiguity between /ɪn/-words and /ɛn/-words for particular speakers, and consequentially, the detections of familiar /ɛn/-words were facilitated for non-merged voices but delayed for merged voices. The debriefing questionnaire has revealed that listeners were mostly unaware of the merger being the focus of the investigation. The dialect-related speaker-adaptation took place swiftly and unconsciously.

The effects of sociolinguistic cues surprisingly surfaced even before listeners experienced clear differences in pronunciations across speakers. It is interesting that listeners seemed to have stronger expectations about the Black speakers' than for the White speakers' pronunciations. Nonetheless, the merged voices led to higher competitions with White than with Black faces in Block 2. This may indicate that listeners were integrating the merged voice better with Black faces than with White faces. Taken with the Block 1 effect, this suggests the listeners from central Ohio associate the /ɪ/ - /ɛ/ merger with African American speech. The lack of effect of dress may indicate that the listeners do not associate the merger with socioeconomic variation or, more likely, that the dress manipulation was not strong enough or specific enough to invoke associations with the merger.

5. REFERENCES

[1] Brown, V. 1990. Phonetic constraints on the merger of /ɪ/ and /ɛ/ before nasals in North Carolina and Tennessee. *The SECOL Review* 14(2), 87-100.

[2] Brown, V.R. 1991. Evolution of the merger of /ɪ/ and /ɛ/ before nasals in Tennessee. *American Speech* 66(3), 303-315.

[3] Conrey, B.P., G.F., Niedzielski, N.A. 2005. Effect of dialect on merger perception: ERP and behavioral correlates. *Brain and Language* 95, 435-449.

[4] Dahan, D., Drucker, S.J., Scarborough, R.A. 2008. Talker adaptation in speech perception: Adjusting the signal or the representations? *Cognition* 108, 710-718.

[5] Goldinger, S.D. 1996. Words and voices: Episodic traces in spoken word identification and recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 22, 1166-1183.

[6] Janson, T., Schulman, R. 1983. Non-distinctive features and their use. *Journal of Linguistics* 19, 321-336.

[7] Johnson, K. 2005. Speaker Normalization in speech perception. In Pisoni, D.B., Remez, R. (eds.), *The Handbook of Speech Perception*. Oxford: Blackwell, 363-389.

[8] Koops, C., Gentry, E., Pantos, A. 2008. The effect of perceived speaker age on the perception of PIN and PEN vowels in Houston, Texas. *University of Pennsylvania Working Papers in Linguistics* 14(2), 93-101.

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

[10] Labov, W., Karan, M., Miller, C. 1991. Near-mergers and the suspension of phonemic contrast. *Language Variation and Change* 3, 33-74.

[11] Massaro, D.W., Cohen, M.M. 1983. Categorical or continuous speech perception: A new test. *Speech Communication* 2, 15-35.

[12] Montgomery, M., Eble, C. 2004. Historical perspectives on the pen/pin merger in southern American English. In Curzan, A., Emmons, K. (eds.), *Studies in the History of the English Language II: Conversations between Past and Present*. Mouton de Gruyter, 429-449

[13] Preston, D. 1997. The south: The touchstone. In Bernstein, C., Nunnally, T., Sabino, R. (eds.), *Language Variety in the South Revisited*. University of Alabama Press.

[14] Strand, E.A. 2000. *Gender Stereotype Effects in Speech Processing.* Ph.D. Dissertation, Ohio State University.

[15] Thomas, E.R. 2001. *An Acoustic Analysis of Vowel Variation in New World English*. American Dialect Society.