Published in final edited form as:

J Phon. 2016 May ; 56: 66–74. doi:10.1016/j.wocn.2016.02.003.

Audiovisual perceptual learning with multiple speakers

Aaron D. Mitchel[a,*], **Chip Gerfen**[b], and **Daniel J. Weiss**[c]

[a]Department of Psychology and Program in Neuroscience, Bucknell University, Lewisburg, PA 17837, USA

[b]Department of World Languages & Cultures, American University, Washington, DC, USA

[c]Department of Psychology and Program in Linguistics, The Pennsylvania State University, University Park, PA, USA

Abstract

One challenge for speech perception is between-speaker variability in the acoustic parameters of speech. For example, the same phoneme (e.g. the vowel in “cat”) may have substantially different acoustic properties when produced by two different speakers and yet the listener must be able to interpret these disparate stimuli as equivalent. Perceptual tuning, the use of contextual information to adjust phonemic representations, may be one mechanism that helps listeners overcome obstacles they face due to this variability during speech perception. Here we test whether visual contextual cues to speaker identity may facilitate the formation and maintenance of distributional representations for individual speakers, allowing listeners to adjust phoneme boundaries in a speaker-specific manner. We familiarized participants to an audiovisual continuum between /aba/ and /ada/. During familiarization, the “b-face” mouthed /aba/ when an ambiguous token was played, while the “D-face" mouthed /ada/. At test, the same ambiguous token was more likely to be identified as /aba/ when paired with a stilled image of the “b-face” than with an image of the “D-face.” This was not the case in the control condition when the two faces were paired equally with the ambiguous token. Together, these results suggest that listeners may form speaker-specific phonemic representations using facial identity cues.

Keywords

Perceptual learning; Multisensory processes; Speech perception; Talker normalization

1. Introduction

A significant obstacle faced by listeners during speech perception is mapping richly variable acoustic information in speech to appropriate categories in the context of a remarkable amount of between-speaker variability in the acoustic parameters of speech. For example, the vowel in *bat* produced by one speaker and the vowel in *bet* produced by a second speaker might have the same basic acoustic structure (Peterson & Barney, 1952). This challenge, known as the *lack of invariance*, is a fundamental property of speech (Liberman, Cooper,

[*]Corresponding author. Tel.: +1 570 577 3890. adm018@bucknell.edu.

Shankweiler, & Studdert-Kennedy, 1967). One mechanism that may contribute to overcoming this challenge is perceptual learning, a process through which listeners adjust their phonetic space in response to the structure of their environmental input (for a review, see Samuel & Kraljic, 2009).

A growing body of research suggests that perceptual learning yields speaker-specific representations of acoustic information (e.g. Eisner & McQueen, 2005; Kraljic & Samuel, 2007; Trude & Brown-Schmidt, 2012). If this correct, it is reasonable to expect that listeners exploit indexical cues to speaker identity to adjust their interpretation of speech. In fact, a particularly strong prediction is that indexical cues to speaker identity alone can induce a change in the boundary of a sound category, even if the acoustic input is held constant. However, to the best of our knowledge, no study has asked whether listeners can dynamically shift their interpretation of identical speech sounds based solely on concurrent indexical cues – in this case, visual cues to speaker identity. In the present study, we adapt a visually-guided perceptual learning paradigm (see Bertelson, Vroomen, & de Gelder, 2003), incorporating multiple speakers during familiarization. If perceptual learning is speaker-specific, we expect the speaker's face provides a context to guide learning by shaping listeners' interpretation of the speech signal.

1.1. Perceptual learning

As noted above, perceptual learning in this context can be defined as a process by which listeners alter their phonemic boundaries for particular sounds based on the context in which those sounds occur (Samuel & Kraljic, 2009). For example, Norris, McQueen, and Cutler (2003) demonstrated that adult listeners adjust their phonemic categories to match the distribution of sounds in a lexically constrained context. Participants who heard an ambiguous speech sound (between /s/ and /f/) in an /f/ context (e.g. in the word *roof*) during familiarization were more likely to later report that the ambiguous sound was an /f/ than participants who heard the same sound in the context of an /s/ (e.g. *house*). This category re-tuning effect can persist up to 24 h after familiarization (Eisner & McQueen, 2006), indicating that perceptual learning results in a lasting shift in the boundaries between sound categories (see Samuel & Kraljic, 2009). Furthermore, the effects of phonetic retuning can be observed at early perceptual stages (e.g. Trude & Brown-Schmidt, 2012) and are not contingent upon episodic memory (Trude, Duff, & Brown-Schmidt, 2014), reflecting a change in the underlying phonetic representation rather than a response bias acquired during familiarization (Clarke-Davidson, Luce, & Sawusch, 2008; Kleinschmidt & Jaeger, 2015).

Speech perception is a fundamentally multisensory process (see Rosenblum, 2008; Massaro, 1998), and visual input is automatically integrated (Soto-Faraco, Navarra, & Alsius, 2004) with the speech signal to form an audiovisual percept. For example, the McGurk effect occurs when incongruent auditory ("ba") and visual ("ga") input is combined to form a unified audiovisual perception ("da") (McGurk & MacDonald, 1976). These illusions are robust hallmarks of the ubiquity of audiovisual integration, as the effect occurs in the face of explicit instructions to attend to a single modality (Buchan & Munhall, 2011) and when the gender or identity of the face and voice do not match (Green, Kuhl, Meltzoff, & Stevens, 1991). Given the role of vision in speech perception, it is necessary to consider how perceptual learning proceeds in an audiovisual context (see also Mitchel, Christiansen, & Weiss, 2014; Mitchel & Weiss, 2014; Lusk & Mitchel, in press).

The first study to investigate perceptual learning of phoneme categories in audiovisual speech was conducted by Bertelson et al. (2003). In this study, the researchers familiarized participants to an audiovisual speech continuum between /aba/ and /ada/, in which the midpoint of the continuum was ambiguous. When the speech signal was ambiguous, the corresponding lip gesture directed interpretation (e.g. the ambiguous midpoint paired with a bilabial lip gesture would be heard as /aba/). For half of the participants, this ambiguous token was paired with the lip movements corresponding to /aba/, while the other half saw the lip movements corresponding to /ada/. During an audio-only test, participants reported hearing the ambiguous token as /aba/ in the former condition or as /ada/ in the latter. The authors described this shift in perception of the ambiguous token as a recalibration of auditory speech categories induced by the lip gestures during familiarization. Visually guided recalibration is equivalent to lexical retuning in effect size (van Linden & Vroomen, 2007) and similarly supports the simultaneous adaptation of an identical sound to multiple phoneme categories (Keetels, Pecoraro, & Vroomen, 2015).

1.2. Speaker-specific perceptual learning

Following the initial studies on perceptual tuning (Norris et al., 2003; Bertelson et al., 2003), subsequent research has supported the view that perceptual tuning may be speaker-specific. For example, Eisner and McQueen (2005) found that perceptual learning with one speaker did not generalize to a novel speaker at test, and Kraljic and Samuel (2007) demonstrated that listeners can adjust their phonemic representations for multiple speakers concurrently. These studies suggest that the representations formed through perceptual learning are speaker-specific and listeners can adjust their interpretation based upon learned properties of the speaker.

Several recent studies have extended this investigation of speaker-specific perceptual learning to the visual domain. Trude and Brown-Schmidt (2012) tested this possibility with a visual-world eye-tracking paradigm in which a target word (e.g. *bake*) was presented with a foil (e.g. *bag*). One of the speakers (male) had a regional accent in which in which *bag* is pronounced /beɪg/; thus, the target and foil would be phonological competitors in this accent, and the foil should momentarily distract the participant away from the target. The other speaker (female) did not have this accent, and so the target and foils were not phonological competitors. Their results revealed significantly greater fixations toward the foil for the accented male speaker than for the unaccented female speaker, indicating that indexical cues (gender of voice or face) influenced the interpretation of the speech signal. Furthermore, van der Zande, Jesse, and Cutler (2014) found that although visually-guided perceptual learning (similar to Bertelson et al., 2003) would generalize to a novel speaker at test, the magnitude of recalibration was greater when tested with the exposure speaker, supporting the notion that visually-guided perceptual learning may also be speaker-specific.

Speaker-specific perceptual learning is well-captured within the ideal adapter framework recently proposed by Kleinschmidt and Jaeger (2015). This framework adopts a Bayesian approach to model how a listener might account for the lack of invariance in speech. The authors propose that listeners update beliefs about the intended output of a speaker based on distributional representations of an individual speaker's acoustic cues. For example, in a study by Newman, Clouse, and Burnham (2001), the distribution of frication centroids for /ʃ/ and /s/ centered around 5400 hz and 5800 hz, respectively, for speaker KSK and around 5000 hz and 5400 hz for another speaker IAF. In an ideal adapter framework, knowledge about each speaker's distribution of acoustic cues would influence the likelihood that a sound belonged to a particular category, guiding the listener to correctly categorize a sound with a frication centroid around 5400 hz as /ʃ/ for speaker A and /s/ for speaker B. This framework therefore predicts that in a perceptual learning task with multiple novel speakers, participants should track and maintain separate phonetic distributions for each speaker and use these distributions to interpret an ambiguous speech signal.

1.3. Present study

Despite the growing evidence that perceptual learning results in speaker-specific representations, no study, to the best of our knowledge, has investigated whether listeners are able to flexibly change their interpretation of an identical speech sound based on fine-grained distinctions in phoneme boundaries established for individual speakers. To date, research in this field has investigated tuning as a function of regional dialect (Trude & Brown-Schmidt, 2012) or one-to-one mapping between gender and a specific interpretation (Kraljic & Samuel, 2007). However, as noted earlier, between-speaker variability results in scenarios where the same or highly similar acoustic information can be interpreted as pertaining to different phonemic categories depending on an individual speaker's distribution of speech productions (see Kleinschmidt & Jaeger, 2015). In the present study, we extend previous work by examining perceptual tuning at the level of the phoneme boundaries for individual speakers. Specifically, we test the ability of listeners to track separate phonetic distributions for multiple speakers and then use this distributional information to adjust their interpretation of an ambiguous speech token.

To investigate these questions, we adapt the visually-guided perceptual learning procedure developed by Bertelson et al. (2003) to include multiple speakers during familiarization, with the goal of determining whether visual indexical cues (i.e. facial identity) can guide audiovisual recalibration of phonetic categories for multiple speakers We present participants with one of two possible familiarization streams, each containing two speakers producing syllables from a continuum between /aba/ and /ada/, the midpoint of which is ambiguous and could be interpreted as either /b/ or /d/. During familiarization, the lip gestures in the dynamic visual display should direct the interpretation of this ambiguous token, bearing some similarity to variants of the McGurk illusion where an unambiguous lip gesture overrides an ambiguous auditory syllable (e.g. Rosenblum, Yakel, & Green, 2000). For example, if the actor synchronously produces a "b" lip gesture with the ambiguous speech sound, then the participant will likely perceive /aba/ (Bertelson et al., 2003).

The first familiarization stream (*biased* condition) uses a bimodal distribution of speaker-to-sound pairings to create distinct phoneme category boundaries for each speaker. This should bias interpretation of the midpoint of the continuum as /b/ for one speaker and /d/ for the other. The second, control familiarization stream (*unbiased* condition), has an even distribution of speaker–sound pairings that should not bias the interpretation of the ambiguous sound in either direction. Following familiarization, participants are tested on phoneme identification while a still image of the B-face or D-face is on the screen. If participants use visual indexical cues to adjust phonemic boundaries, then in the biased condition, phoneme categorization should be influenced by the presence of the speaker's face such that their perception of ambiguous tokens follows the distributional information available in the visual display. In contrast, there should be no difference in categorization of test items in the unbiased condition. The present study therefore provides a stringent test case for speaker-specific perceptual learning. The audio continuum is identical for both speakers and the *only* contextual cue to adjust phonetic boundary is a still image of the speaker's face. Consequently, evidence of different phonetic boundaries for the two speakers would provide a robust demonstration of speaker-specificity during audiovisual recalibration of speech.

2. Method

2.1. Participants

Sixty-seven monolingual English speakers (44 female, 23 male) from the Pennsylvania State University participated in the experiment for course credit. An additional participant's data were excluded from analysis due to equipment failure during the familiarization phase. 35 participants (23 female, 12 male) were randomly assigned to the experimental condition (biased distribution) and 32 participants (21 female, 11 male) were assigned to the control condition (unbiased distribution). All participants gave informed consent.

2.2. Stimuli

The stimuli in the present study consisted of a seven-point speech continuum between /aba/ and /ada/ that was then synced with videos of two different speakers making [b] or [d] visemic lip gestures (i.e. visual cues to phonetic content).

2.2.1. Audio continuum—The audio stimuli were modeled after the stimuli in Bertelson et al. (2003). We created a 7-point continuum between /aba/ and /ada/ by first recording a male native English speaker producing /aba/. The total duration of the recording was 640 ms with a stop closure of ~167 ms. This recording was then synthesized using Akustyk (Plichta, 2010) along with Praat (Boersma & Weenink, 2011). We varied the locus of the second formant (F2) by approximately 66 hz intervals to create a seven point continuum in the F2 transition (50 ms before and after the stop consonant) that signals place of articulation (see Fig. 1).

2.2.2. Visual stimuli—Using a Sony Handycam mounted on a tripod, we recorded two separate actors lip-syncing "aba" and "ada." The videos (4 in total) were imported into Adobe Premiere® and were hand-edited to remove the original audio. Each of these four lip movement videos (both actors producing both lip movements) were combined and synchronized with each of the 7 audio stimuli in the continuum between /aba/ and /ada/. This resulted in 28 individual video clips.

Based on the findings of Bertelson et al. (2003), we expected the lip gesture to direct the interpretation of the ambiguous speech signal. For example, the third token in the continuum was more likely to be perceived as /aba/ in the absence of any visual input. When paired with an "ada" lip gesture, participants should be more likely to perceive this same token as /ada/. Thus, while the auditory stimuli remain invariant, the visual context should systematically shift the location of the phoneme boundary.

2.2.3. Familiarization streams—We combined the 28 individual video clips to create two familiarization streams. The first, *unbiased* familiarization stream served as our baseline condition. Each of the 28 video clips was included in the stream 5 times, resulting in a combined familiarization stream of 140 video clips. The sequence order for the clips was randomly generated and then, using Adobe Premiere®, the clips were combined into a Quicktime movie of 2 min 48 s.

In the *biased* familiarization stream, we manipulated the frequency distributions of the video clips to bias the interpretation of the ambiguous token (i.e., token 4) depending on which of the two speakers was producing the token. That is, for the first actor (hereafter referred to as the B face), tokens 1–4 in the /aba/-/ada/ continuum were more often paired with an "aba" lip gesture, whereas tokens 6–7 were more likely to be paired with an "ada" lip gesture and token 5 was paired equally. The frequency distribution of these pairings should provide a cue to the listener that, for this speaker, the phoneme boundary between /aba/ and /ada/ should be centered closer to the 5th token in the continuum; thus, token 4 should be interpreted as /aba/. In contrast, for the second actor (D face), tokens 4–7 were more often paired with an "ada" lip gesture, tokens 1–2 were more likely to be paired with "aba", and token 3 was equally paired. Thus, for this speaker, the phoneme boundary is centered closer to token 3, and token 4 should be interpreted as /ada/ (see Fig. 2).

Each auditory token was heard 20 times, resulting in a familiarization stream of 140 video clips. The sequence order for the video clips was generated randomly. All 140 video clips were then concatenated in Adobe Premiere to produce a video of 2 min 48 s. This video was exported as a Quicktime (.mov) file.

Critically, each face was paired with both lip gestures an equal number of times (35 each). Participants therefore could not form a simple association between a face and a particular interpretation. It was not the case that the B face was simply more likely to be paired with "aba." The only difference between the two speakers was in the distributional pattern of gesture-token pairings; thus, any effect of speaker on interpretation of the ambiguous token at post-test should be due to participants' ability to use distributional learning to form separate phonemic profiles for each speaker.

2.3. Procedure

The experiment consisted of three phases: pretest, familiarization, and post-test (see Fig. 3). In the pretest, participants were presented with each of the seven items from the auditory continuum and asked to determine if it was "aba" or "ada". Each item was tested five times for a total of 35 trials. Both the pretest and post-test were presented using E-prime software (version 2.0) with Sennheiser HD 280 pro headphones. The order of test trials was randomized for each participant.

During familiarization, participants viewed either the biased or unbiased familiarization stream. The procedures for both familiarization conditions were identical. Participants were instructed to watch a brief movie, after which they would be tested on information picked up from the movie. The familiarization stream was presented with iTunes software through headphones.

Following familiarization, participants completed a two-alternative forced choice test in which participants listened to a speech token and indicated if the speaker was saying “aba” or “ada” by pressing the corresponding key on the keyboard. In addition, while each speech item was played, a static image of one of the actors was presented on screen (see Fig. 3). The static images were still frames from the familiarization video depicting the actors producing vowels. The image appeared 200 ms prior to the onset of the speech item (200 ms SOA) and remained on the screen throughout the duration of the speech item. There was no visemic information that could disambiguate the test file as either /aba/ or /ada/ (see Fig. 3). Thus, the faces at post-test only provided an indexical cue to speaker. Participants were instructed to respond as quickly and as accurately as possible.

Consistent with Bertelson et al. (2003), the test items consisted of the three middle-most points in the auditory continuum (tokens 3–5). We did not test all continuum items because 1) we wished to maintain consistency with previous research and 2) the middle three items provide enough range to detect a shift in phoneme boundary while reducing the risk of retraining at test. The auditory test items were presented 10 times each with both of the faces, resulting in 60 test trials presented in a random order. The same test was used in the biased and unbiased conditions.

3. Results

3.1. Pretest

The pretest revealed a reliable phoneme boundary at approximately the 4th continuum token (see Fig. 4). To ensure that there were no pre-existing, systematic differences in phoneme boundaries for participants in the two familiarization conditions, we compared their pretest performance. Participants in the biased condition identified token 4 as “aba” on 61% of pretest trials (SD=30%); and in the unbiased condition token 4 was also identified as “aba” on 61% of trials (SD=29%). Using a sigmoid function, a curve was fitted to the data in both conditions (see Fig. 4). Based on the results of this curve fitting, the phoneme boundary between /b/ and /d/ was calculated by estimating the token number where the proportion of response was evenly distributed between “aba” and “ada” (i.e. y=.50). In the biased condition, the mean phoneme boundary was 4.17 (95% CI: 4.03 to 4.31). In the unbiased condition, the mean phoneme boundary was 4.13 (95% CI: 4.04 to 4.22). The pretest results establish a baseline response pattern and verify that our continuum stimuli result in a categorical shift in phoneme perception.

3.2. Post-test

We calculated the proportion of “aba” responses for each auditory token (3, 4, 5) across face and familiarization conditions. The means and standard errors are reported in Fig. 5. Separate paired samples *t*-tests (two-tailed) were conducted for all three post-test auditory tokens to compare response profiles for the different face conditions (i.e. B face vs. D face). For the biased condition, participants were significantly more likely (corrected α=.016) to identify the ambiguous token, 4, as “aba” when it was paired with the B face (*M*=0.45, SD=0.23) than when it was paired with the D face (*M*=0.38, SD=0.27): *t*(34)=2.88, *p*=.007, *d*=0.49. In contrast, there was no difference in the proportion of “aba” responses for token 3 (*t*(34)=0.29, *p*=0.777, *d*=0.05) or token 5 (*t*(34)=−0.27, *p*=0.79, *d*=−0.05). This was not surprising, given the steep slope of our categorization function from the pre-test. Since there was little ambiguity in the perception of tokens 3 and 5, an effect of retuning on either of these auditory items was unlikely. For the unbiased condition, there were no significant differences between the face conditions in the proportion of “aba” responses for any of the auditory tokens: 3, *t*(31)=0.00, *p*=1.00, *d*=0.00; 4, *t*(31)=0.11, *p*=.91, *d*=0.02; 5, *t*(31)=0.84, *p*=.407, *d*=0.15.

To compare the effect of face display on interpretation of the ambiguous item (token 4) across familiarization conditions, we conducted a 2 (face) × 2 (familiarization condition) repeated measures ANOVA. In particular, we were interested in whether the effect of face was greater in the biased condition relative to the unbiased condition – if there were a greater effect in the unbiased condition, then this would be equivalent to a null finding. Given this directional hypothesis, we used Howell’s (1997) procedure for testing a one-tailed, directional interaction term. This ANOVA revealed a significant within-subjects effect (two-tailed) for face [F(1, 65)=4.08, *p*=.048, $\eta_p^2=.059$], no significant between-subjects effect (two-tailed) for familiarization condition [F(1, 65)=0.06, *p*=.803, $\eta_p^2=.001$], and a significant one-tailed interaction between face and familiarization condition [F(1, 65)=3.42, *p*=.035, $\eta_p^2=.050$].

Complementing this ANOVA analysis, we also performed a direct comparison of effect sizes. This statistical comparison, detailed in Rosenthal and Rosnow (1991), tests the hypothesis that the effect of face was greater in the biased than unbiased condition by comparing the effect size of face presentation (B face–D face) between familiarization conditions (see Fig. 6). Cohen’s *d* for the face effect was 0.49 (*r*=.443, Fisher’s *z*=.48) in the biased condition and 0.02 (*r*=.020, Fisher’s *z*=.02) in the unbiased condition. There was a significantly larger effect size of face presentation in the biased condition than in the unbiased condition: Z=1.78, *p*(one-tailed)=.038. This provides additional evidence that the face of the speaker influenced interpretation of the ambiguous token in the biased condition, but not in the unbiased condition.

It is worth noting that although the retuning effect observed in the biased condition (a 7% shift) is both reliable and comparable to other perceptual learning studies using stop consonants (e.g. Kraljic & Samuel, 2006, 2007), the effect is smaller than what has been observed in some visually-guided perceptual learning studies (e.g. Bertelson et al., 2003; van der Zande et al., 2014). However, as we noted above, the present study was a stringent test case for speaker-specific learning. The audio stimuli were identical and therefore the only signal for listeners to shift their category boundary was a static image of the speaker; in light of these methodological differences, the observed effect in the biased condition provides compelling evidence for speaker-specific learning. Moreover, the previous visual recalibration studies utilized a single speaker design; thus, the exposure phase provided a consistent direction of adaptation (i.e. the ambiguous item was consistently biased toward either /b/ or /d/), whereas in our distributional design, the direction of retuning was contingent upon the speaker. Finally, we observed a retuning effect despite contextual factors that might bias listeners away from recalibration. For example, participants viewed two different faces while hearing only a single voice and may have been less prone to phonetic retuning in this narrow context.

4. Discussion

4.1. Summary

We presented participants with one of two familiarization streams containing two speakers' lip-synced tokens from a single speech continuum. The biased stream contained distributional cues that the two speakers had distinct phonemic boundaries, such that for one face (B face) an ambiguous speech sound should be interpreted as “aba” and for the other face (D face) the same sound should be interpreted as “ada.” Consistent with our prediction, participants shifted their phoneme boundaries based on the face presented at test, as they were significantly more likely to identify the ambiguous token as “aba” when paired with the B face than when paired the D face. In contrast, when there were no distributional cues that the speakers had differing phonemic boundaries (unbiased condition), participants were equally likely to interpret the ambiguous speech sound as either phoneme regardless of the accompanying face. Furthermore, the effect of faces was significantly larger in the biased condition than in the unbiased condition. This suggests that the identity of the speaker, derived from indexical information in the facial image, triggered participants to adjust their phonemic boundaries in a speaker-specific manner. This required participants to recognize, from only brief familiarization, that each speaker had a unique distribution of speech sounds along the /aba-ada/ continuum and subsequently to use this information to dynamically adjust his interpretation of an ambiguous speech sound.

4.2. Implications for perceptual learning

As noted in the Section 1, there is a body of evidence suggesting that perceptual learning of speech can result in speaker-specific representations (Kraljic & Samuel, 2007; Trude & Brown-Schmidt, 2012; Trude et al., 2014; see also Kleinschmidt & Jaeger, 2015). Our results build on this growing literature by providing the first demonstration that learners can dynamically shift their interpretation of an identical speech sound based on visual indexical information (i.e. the speaker's face). Our findings also extend work on visually-induced perceptual learning of speech (e.g. Bertelson et al., 2003; van Linden & Vroomen, 2007; van der Zande et al., 2014). Previous research established that participants can use visemic information to not only adjust their interpretation of ambiguous speech sounds but to retune (or *recalibrate*) their phonemic representations based on this visual context. The present study extends research in this field by demonstrating that visual cues can support perceptual learning with multiple speakers within the same test, providing evidence for speaker-specific learning. In addition, the present study illustrates the role that indexical information plays in speech and language processes (see also Mitchel & Weiss, 2010; Mani & Schneider, 2013). The results of the present study, in which facial indexical cues activated specific distributions of phonemic information, provide support for a multimodal representation of talker identity (Campanella & Belin, 2007), consistent with studies demonstrating the integration of visual and auditory indexical codes during speech perception (Mullennix & Pisoni, 1990; Von Kriegstein, Kleinschmidt, Sterzer, & Giraud, 2005; Von Kriegstein & Giraud, 2006).

It is worth noting that previous research on visually-guided perceptual learning (e.g. Bertelson et al., 2003; van der Zande et al., 2014) has occasionally employed both a recalibration paradigm (as was used in the present study) and a selective adaptation paradigm, in which exposure to a unambiguous auditory token paired with a visual gesture results in a decreased tendency to perceive that auditory token at test (i.e. exposure to a congruent audiovisual /b/ reduces the categorization of an ambiguous sound as /b/). The design of the present study precluded the use of an adaptation procedure, since it was essential to balance the frequency of lip gesture and speech token pairings across actors. However, future work that is not predicated upon distributional learning may be able to examine whether selective adaptation with multiple speakers similarly exhibits speaker-specific learning, and this may provide additional insight into the stage or stages of speech processing at which speaker-specific audiovisual recalibration occurs. For example, in contrast to token 4, we did not observe a perceptual shift for tokens 3 and 5. This lack of recalibration may reflect the relative perceptual distance between tokens on the continuum—if tokens 3 and 5 were well within the /b/ or /d/ category (each was largely uniform in interpretation at pre-test), it would be difficult to induce a significant shift in perception and therefore recalibration would likely be localized to tokens at or near the category boundary (see Kraljic & Samuel, 2007). Selective adaptation, through repeated exposure, may be able to induce a broader categorical shift, which would provide evidence for learning at the perceptual level rather than the decisional (i.e. response criterion) level.

It is also worth noting that perceptual learning does not appear to be uniformly speaker-specific. Several studies have also found evidence of generalization of learning in some contexts and the absence of generalization in others. For example, Kraljic and Samuel (2007) found evidence of generalization across speakers when using a VOT continuum (/d-t/; see also Kraljic & Samuel, 2006), but found no evidence of generalization when the stimuli consisted of fricatives (/s–ʃ/; see also Eisner & McQueen, 2005). The authors proposed that the reason for generalization in the former context, but not the latter, was that differences in VOTs do not vary predictably between speakers in the same way that a spectral shift (such as in a fricative continuum) more reliably denotes a change in speaker. Based on these results, both Kraljic and Samuel (2007) and Kleinschmidt and Jaeger (2015) argue that the perceptual learning mechanism is flexible, supporting generalization when appropriate while fostering distinct speaker-specific representations in other contexts.

Likewise, van der Zande et al. (2014) recently found evidence of generalization in a visually-guided perceptual learning study. After familiarization to a single speaker producing ambiguous syllables in a constrained visual context – either /b/ or /d/ lip gestures – participants shifted their interpretation of an auditory test item (exhibiting perceptual learning) produced by either the exposure speaker or a novel speaker. On the surface, such generalization appears to be inconsistent with speaker-specific retuning, as evidenced in the present study. However, contextual cues may have led to generalization in van der Zande and encapsulation in the present study. In the present study, the presence of multiple talkers may have signaled a greater likelihood of multiple distributions in the input and prompted listeners to encapsulate learning in a speaker specific manner, whereas in van der Zande et al. (2014) the presence of only a single talker provided no such prompt. Future studies will examine the issue of generalization by testing whether perceptual learning with multiple speakers can be generalized to novel speakers at test.

4.3. Models of talker adaptation

The results of the present study may also inform models of talker adaptation during speech perception. Talker adaptation is the process of adjusting one’s phonetic space to accommodate for between-speaker variation in the production of speech sounds. Many different mechanistic models have been proposed to account for talker adaptation, including vocal tract normalization (e.g. Rand, 1971), exemplar-based theories (see Johnson, 2008, for a review), and Kleinschmidt and Jaeger’s (2015) ideal adapter framework, all of which propose that listeners make use of visual information during talker adaptation. Vocal tract normalization proposes that listeners make quick inferences about the size of the speaker’s vocal tract, which then allows them to adjust their interpretation of the speech signal. However, since there was no adaptation observed in the unbiased condition, which contained the same visual cues to vocal tract length as the biased condition, the results of the present study do not lend support for vocal tract normalization.

Exemplar theories, in contrast, maintain that talker adaptation occurs by basing speech interpretation on a set of experienced instances stored for each stimulus and each speaker (Johnson, 1997). These exemplars structure speaker-specific representations that serve as a reference point in decoding the speech signal. Perceptual learning may provide a mechanism to support exemplar-based talker adaptation, accounting for classic hallmarks of adaptation (see Johnson, 2008) such as context dependency (e.g. Ladefoged & Broadbent, 1957), domain specificity (e.g. Nygaard & Pisoni, 1998), and the presence of cross-modal interactions (e.g. Johnson, Strand & D’Imperio, 1999), though exemplar-based models are unable to account for generalization of perceptual learning to novel speakers (e.g. Kraljic & Samuel, 2007). Future work can further elucidate the role of perceptual learning in talker adaptation by examining these effects and their sensitivity to distributional properties of the input.

Finally, the findings from the present study are also consistent with an ideal adapter framework (Kleinschmidt & Jaeger, 2015) account of talker adaptation. As mentioned above, this is a computational model wherein listeners form beliefs about the talker’s generative model (i.e. the distribution of acoustic cues that manifest as a particular linguistic unit, such as a phonetic category) that are updated based on knowledge of the talker. Unlike exemplar-based theories, the ideal adapter framework operates on abstract distributional information, rather than specific episodic events. Therefore, a strength of this framework is its ability to account for both talker-specificity and talker-generalization during phonetic recalibration. From this perspective, it may be possible to adapt the paradigm developed here to explore how listeners decide when it is appropriate to generalize knowledge to a novel speaker. For instance, Kleinschmidt and Jaeger (2015) propose that listeners should generalize a generative model to a novel speaker that is similar (either in distribution of acoustic cues or in social group) to a previously encountered speaker. Thus, it is reasonable to predict that listeners might generalize to speakers who are visually similar (e.g. faces that overlap in a multidimensional face space; Valentine, 1991). Future work can therefore test the predictions of an ideal adapter framework by investigating whether visually-guided perceptual learning will generalize at test to novel speakers with similar and dissimilar facial features and dimensions.

References

Bertelson P, Vroomen J, de Gelder B. Visual recalibration of auditory speech identification: a McGurk aftereffect. Psychological Science. 2003; 14:592–597. [PubMed: 14629691]

Boersma, P., Weenink, D. Praat: doing phonetics by computer (Version 5.2.11), [Computer Software]. 2011. Retrieved January 18, 2011 from ⟨http://www.praat.org/⟩

Buchan JN, Munhall KG. The influence of selective attention to auditory and visual speech on the integration of audiovisual speech information. Perception. 2011; 40:1164–1182. [PubMed: 22308887]

Campanella S, Belin P. Integrating face and voice in person perception. Trends in Cognitive Sciences. 2007; 11:535–543. [PubMed: 17997124]

Clarke-Davidson CM, Luce PA, Sawusch JR. Does perceptual learning in speech reflect changes in phonetic category representation or decision bias? Perception psychophysics. 2008; 70:604–618. [PubMed: 18556922]

Eisner F, McQueen JM. The specificity of perceptual learning in speech processing. Perception Psychophysics. 2005; 67:224–238. [PubMed: 15971687]

Eisner F, McQueen JM. Perceptual learning in speech: stability over time. Journal of the Acoustical Society of America. 2006; 119:1950–1953. [PubMed: 16642808]

Green KP, Kuhl PK, Meltzoff AN, Stevens EB. Integrating speech information across talkers, gender, and sensory modality: female faces and male voices in the McGurk effect. Perception psychophysics. 1991; 50:524–536. [PubMed: 1780200]

Howell, DC. Statistical methods for psychology. 4th. Belmont, CA: Duxbury Press; 1997.

Johnson, K. Speech perception without speaker normalization: an exemplar model. In: Johnson, K., Mullennix, J., editors. Talker variability in speech processing. New York: Academic Press; 1997. p. 145-166.

Johnson, K. Speaker normalization in speech perception. In: Pisoni, DB., Remez, RE., editors. The handbook of speech perception. Oxford, UK: Blackwell Publishing Ltd; 2008. p. 363-389.

Johnson K, Strand EA, D'Imperio M. Auditory-visual integration of talker gender in vowel perception. Journal of Phonetics. 1999; 27:359–384.

Keetels M, Pecoraro M, Vroomen J. Recalibration of auditory phonemes by lipread speech is ear-specific. Cognition. 2015; 141:121–126. [PubMed: 25981732]

Kleinschmidt DF, Jaeger TF. Robust speech perception: recognize the familiar, generalize to the similar, and adapt to the novel. Psychological Review. 2015; 122:148. [PubMed: 25844873]

Kraljic T, Samuel AG. Generalization in perceptual learning for speech. Psychonomic Bulletin and Review. 2006; 13:262–268. [PubMed: 16892992]

Kraljic T, Samuel AG. Perceptual adjustments to multiple speakers. Journal of Memory and Language. 2007; 56:1–15.

Ladefoged P, Broadbent DE. Information conveyed by vowels. The Journal of the Acoustical Society of America. 1957; 29:98–104.

Liberman AM, Cooper FS, Shankweiler DP, Studdert-Kennedy M. Perception of the speech code. Psychological Review. 1967; 74:431. [PubMed: 4170865]

Lusk LG, Mitchel AD. Differential gaze patterns on eyes and mouth during audiovisual speech segmentation. Frontiers in Psychology. 2016 (in press).

Mani N, Schneider S. Speaker identity supports phonetic category learning. Journal of Experimental Psychology: Human Perception and Performance. 2013; 39:623–629. [PubMed: 23148468]

Massaro, DW. Perceiving talking faces: from speech perception to a behavioral principle. Cambridge, MA: MIT Press; 1998.

McGurk H, MacDonald J. Hearing lips and seeing voices. Nature. 1976; 264:746–748. [PubMed: 1012311]

Mitchel AD, Christiansen MH, Weiss DJ. Multimodal integration in statistical learning: evidence from the McGurk illusion. Frontiers in Psychology. 2014; 5(407):1–6. [PubMed: 24474945]

Mitchel AD, Weiss DJ. What's in a face? Visual contributions to speech segmentation. Language and Cognitive Processes. 2010; 25:456–482.

Mitchel AD, Weiss DJ. Visual speech segmentation: using facial cues to locate word boundaries in continuous speech. Language, Cognition Neuroscience. 2014; 29:771–780.

Mullennix JW, Pisoni DB. Stimulus variability and processing dependencies in speech perception. Perception Psychophysics. 1990; 47:379–390. [PubMed: 2345691]

Newman RS, Clouse SA, Burnham JL. The perceptual consequences of within-talker variability in fricative production. The Journal of the Acoustical Society of America. 2001; 109:1181–1196. [PubMed: 11303932]

Norris D, McQueen JM, Cutler A. Perceptual learning in speech. Cognitive Psychology. 2003; 47:204–238. [PubMed: 12948518]

Nygaard LC, Pisoni DB. Talker-specific learning in speech perception. Perception Psychophysics. 1998; 60:355–376. [PubMed: 9599989]

Peterson GE, Barney HL. Control methods used in the study of vowels. Journal of the Acoustical Society of America. 1952; 24:175–184.

Plichta, B. Akustyk for Praat [computer software]. 2010. Available from 〈http://bartus.org/akustyk/index.php〉

Rand TC. Vocal tract size normalization in the perception of stop consonants. The Journal of the Acoustical Society of America. 1971; 50:139.

Rosenblum LD. Speech perception as a multimodal phenomenon. Current Directions in Psychological Science. 2008; 17:405–409. [PubMed: 23914077]

Rosenblum LD, Yakel DA, Green KP. Face and mouth inversion effects on visual and audiovisual speech perception. Journal of Experimental Psychology: Human Perception and Performance. 2000; 26:806–819. [PubMed: 10811177]

Rosenthal, R., Rosnow, RL. Essentials of behavioral research: Methods and data analysis. 2nd. Boston, MA: McGraw Hill; 1991.

Samuel AG, Kraljic T. Perceptual learning for speech. Attention, Perception, Psychophysics. 2009; 71:1207–1218.

Soto-Faraco S, Navarra J, Alsius A. Assessing automaticity in audiovisual speech integration: evidence from the speeded classification task. Cognition. 2004; 92:B13–B23. [PubMed: 15019556]

Trude AM, Brown-Schmidt S. Talker-specific perceptual adaptation during online speech perception. Language and Cognitive Processes. 2012; 27:979–1001.

Trude AM, Duff MC, Brown-Schmidt S. Talker-specific learning in amnesia: insight into mechanisms of adaptive speech perception. Cortex. 2014; 54:117–123. [PubMed: 24657480]

Valentine T. A unified account of the effects of distinctiveness, inversion, and race in face recognition. The Quarterly Journal of Experimental Psychology. 1991; 43:161–204. [PubMed: 1866456]

van der Zande P, Jesse A, Cutler A. Cross-speaker generalisation in two phoneme-level perceptual adaptation processes. Journal of Phonetics. 2014; 43:38–46.

van Linden S, Vroomen J. Recalibration of phonetic categories by lipread speech versus lexical information. Journal of Experimental Psychology: Human Perception and Performance. 2007; 33:1483. [PubMed: 18085958]

Von Kriegstein K, Giraud AL. Implicit multisensory associations influence voice recognition. PLoS: Biology. 2006; 4(10):e326. [PubMed: 17002519]

Von Kriegstein K, Kleinschmidt A, Sterzer P, Giraud AL. Interaction of face and voice areas during speaker recognition. Journal of Cognitive Neuroscience. 2005; 17:367–376. [PubMed: 15813998]

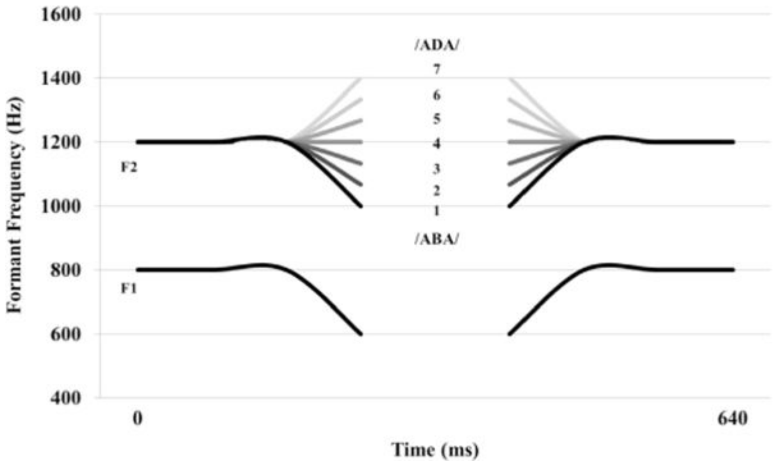


Fig. 1.

First (F1) and second (F2) formant transitions for the seven tokens from the audio continuum, ranging from the initial /aba/ token (1), the midpoint token (4), and the final /ada/ token (7).

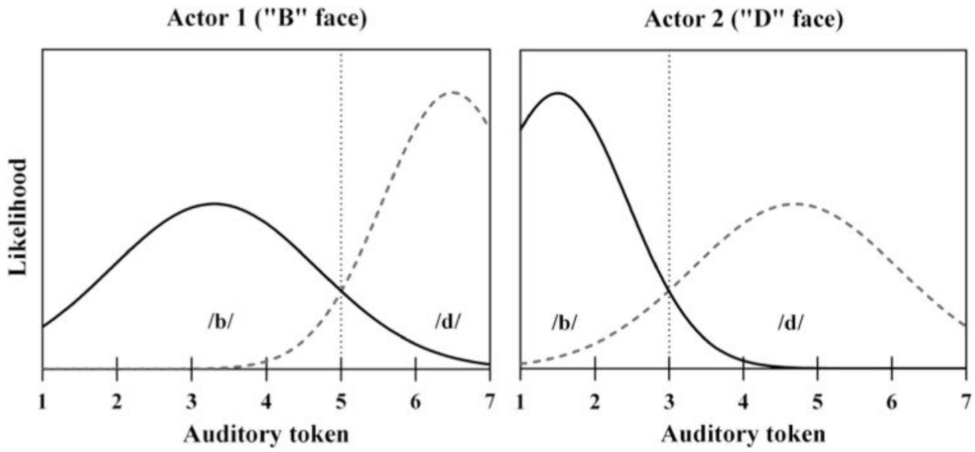
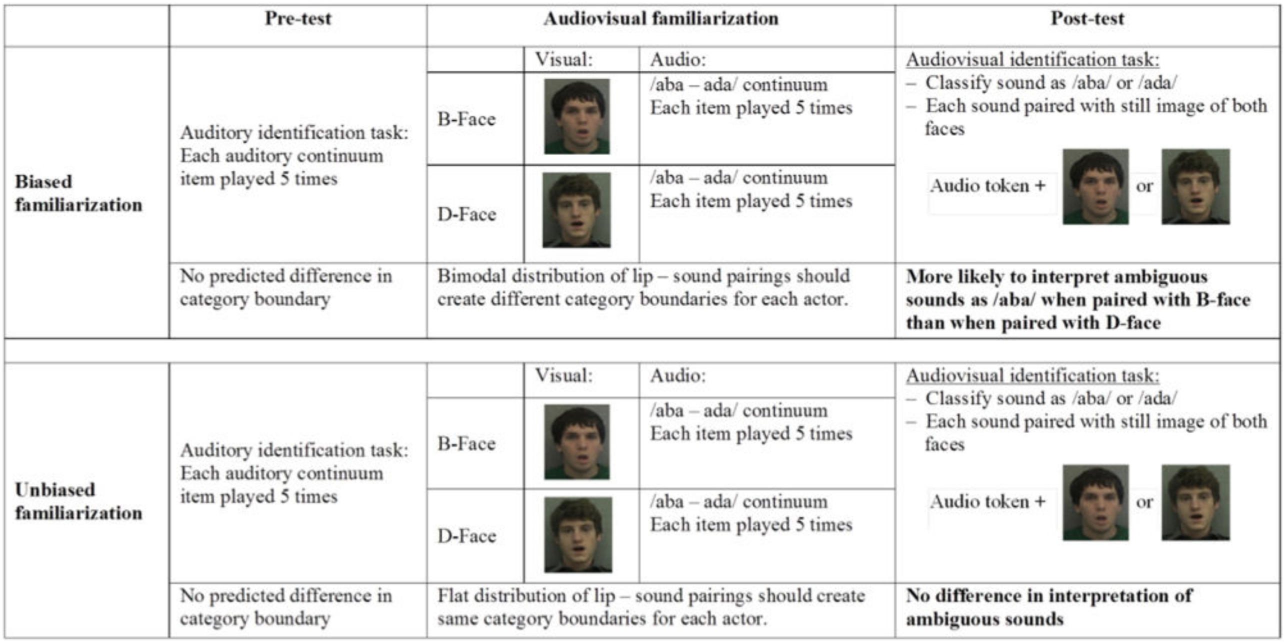


Fig. 2.

The distribution of pairings between lip gestures and tokens from the auditory continuum (and probable categorization) for each actor in the *biased* familiarization. Solid lines represent an “aba” lip gesture, whereas the dashed lines represent an “ada” lip gesture. The dotted vertical line in each frame represents a hypothetical category boundary on the /b–d/ continuum for each speaker.



	Pre-test	Audiovisual familiarization			Post-test
			Visual:	Audio:	Audiovisual identification task: – Classify sound as /aba/ or /ada/ – Each sound paired with still image of both faces Audio token + or
Biased familiarization	Auditory identification task: Each auditory continuum item played 5 times	B-Face		/aba – ada/ continuum Each item played 5 times	
		D-Face		/aba – ada/ continuum Each item played 5 times	
	No predicted difference in category boundary	Bimodal distribution of lip – sound pairings should create different category boundaries for each actor.			**More likely to interpret ambiguous sounds as /aba/ when paired with B-face than when paired with D-face**
			Visual:	Audio:	Audiovisual identification task: – Classify sound as /aba/ or /ada/ – Each sound paired with still image of both faces Audio token + or
Unbiased familiarization	Auditory identification task: Each auditory continuum item played 5 times	B-Face		/aba – ada/ continuum Each item played 5 times	
		D-Face		/aba – ada/ continuum Each item played 5 times	
	No predicted difference in category boundary	Flat distribution of lip – sound pairings should create same category boundaries for each actor.			**No difference in interpretation of ambiguous sounds**

Fig. 3.

Schematic of procedure and predictions for the biased and unbiased conditions.

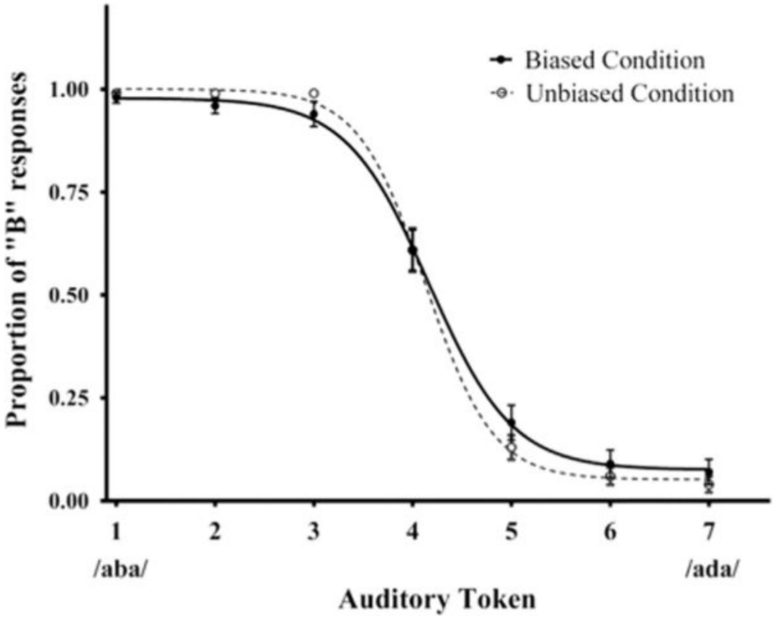


Fig. 4.

Mean proportion of responses labeled “aba” for each token in the auditory continuum in the pretest for participants who subsequently viewed the unbiased or biased familiarization stream. Note that there was no procedural difference between the biased and unbiased conditions for the pre-test.

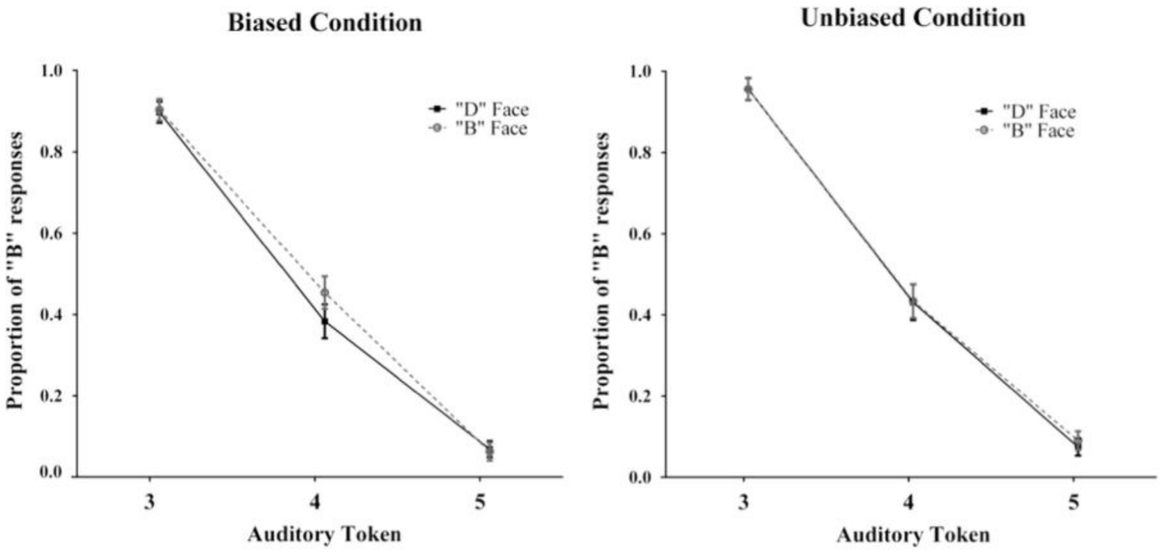


Fig. 5.

Mean proportion of responses labeled “aba” for the three test items in the post-test for both the biased and unbiased conditions. Lines are separated by which face was presented at test. Note: *A?−1, A?*, and *A?+1* refer to tokens 3, 4, and 5, respectively, from the auditory continuum. Error bars represent ±SEM.

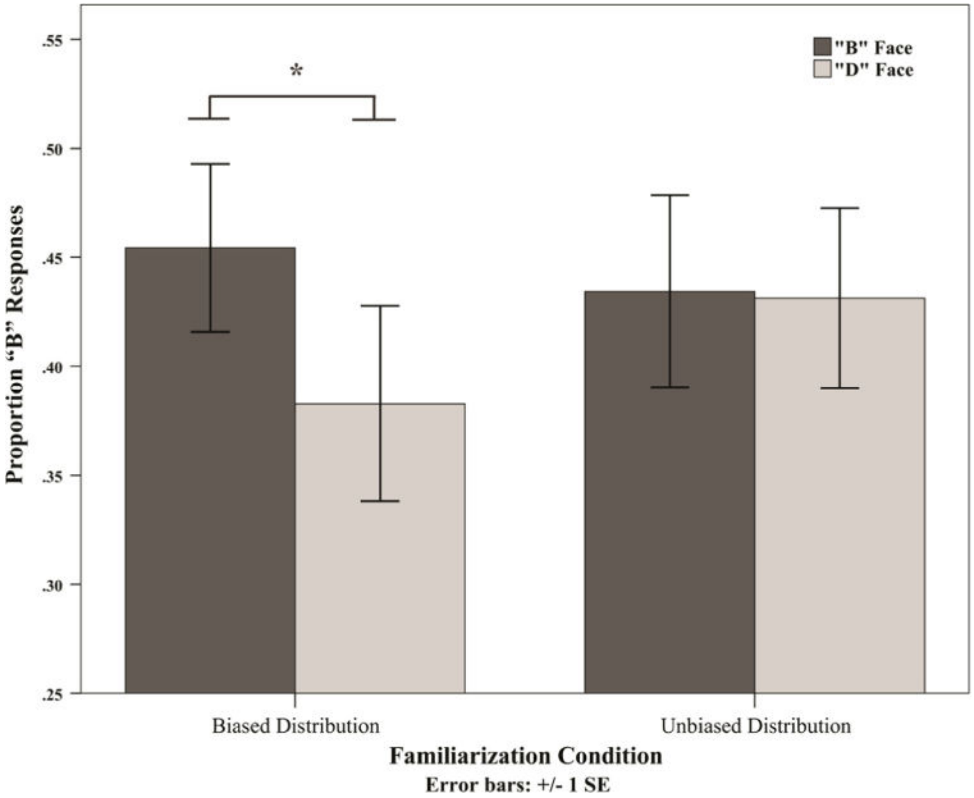


Fig. 6.
Mean proportion of responses labeled "aba" for the ambiguous midpoint (token 4) of the auditory continuum.