Journal of Phonetics 56 (2016) 66–74

Contents lists available at ScienceDirect

Journal of Phonetics

journal homepage: www.elsevier.com/locate/phonetics

Research Article

Audiovisual perceptual learning with multiple speakers

Aaron D. Mitchel [a,*], Chip Gerfen [b], 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

ARTICLE INFO

Article history:
Received 21 August 2015
Received in revised form
18 January 2016
Accepted 25 February 2016

Keywords:
Perceptual learning
Multisensory processes
Speech perception
Talker normalization

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.

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, 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

* Corresponding author. Tel.: +1 570 577 3890.
E-mail address: adm018@bucknell.edu (A.D. Mitchel).

0095-4470/$ - see front matter © 2016 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.wocn.2016.02.003

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