Journal of Memory and Language 69 (2013) 527–545

Contents lists available at SciVerse ScienceDirect

Journal of Memory and Language

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

No delays in application of perceptual learning in speech recognition: Evidence from eye tracking

Holger Mitterer [a,*], Eva Reinisch [b]

[a] Max Planck Institute for Psycholinguistics, The Netherlands
[b] Institute of Phonetics and Speech Processing, University of Munich, Germany

ARTICLE INFO

Article history:
Received 4 October 2012
revision received 2 July 2013
Available online 2 August 2013

Keywords:
Speech perception
Perceptual learning
Eye tracking

ABSTRACT

Three eye-tracking experiments tested at what processing stage lexically-guided retuning of a fricative contrast affects perception. One group of participants heard an ambiguous fricative between /s/ and /f/ replace /s/ in s-final words, the other group heard the same ambiguous fricative replacing /f/ in f-final words. In a test phase, both groups of participants heard a range of ambiguous fricatives at the end of Dutch minimal pairs (e.g., *roos-roof*, 'rose'-'robbery'). Participants who heard the ambiguous fricative replacing /f/ during exposure chose at test the f-final words more often than the other participants. During this test-phase, eye-tracking data showed that the effect of exposure exerted itself as soon as it could possibly have occurred, 200 ms after the onset of the fricative. This was at the same time as the onset of the effect of the fricative itself, showing that the perception of the fricative is changed by perceptual learning at an early level. Results converged in a time-window analysis and a Jackknife procedure testing the time at which effects reached a given proportion of their maxima. This indicates that perceptual learning affects early stages of speech processing, and supports the conclusion that perceptual learning is indeed perceptual rather than post-perceptual.

Introduction

Even though listeners become attuned to the typical pronunciations of the sounds of their native language during the first year of life (Werker & Tees, 1984), recent evidence shows that these established phonetic categories remain surprisingly flexible (for a review, see Samuel & Kraljic, 2009). This flexibility can be experienced in everyday life when listening to speakers with different regional and foreign accents: as we become more familiar with their pronunciation peculiarities, their speech becomes easier to understand. This has been demonstrated empirically on a global level as good recognition of foreign-accented words after some exposure (Bradlow & Bent, 2008) but also on a more fine-grained phonemic level as listeners adjust to the unusual pronunciation of a particular native-language segment (starting with the seminal study by Norris, McQueen, & Cutler, 2003). Even though it seems well established (anecdotally and empirically) that adjustment to a speaker does occur, what has not been addressed so far is when during speech processing the new knowledge about the pronunciation of a segment is applied. Once we know that a certain speaker produces a phoneme in an unusual fashion, do we immediately interpret new instances of this phoneme in relation to our prior experience? Or is early phonetic processing not affected by perceptual learning, and only the final decision about the segment's identity is influenced by the newly learned knowledge? (Similar to the conceptualization of auditory and visual processing in Massaro's, 1998, FLMP model.) The present study set out to address this question by revealing the cognitive stages of speech processing at which knowledge about pronunciation variants is taken into account. Specifically we asked whether retuned

* Corresponding author. Address: University of Malta, Department of Cognitive Science, Msida, Malta.
E-mail address: holger.mitterer@um.edu.mt (H. Mitterer).

0749-596X/$ - see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.jml.2013.07.002

phonetic categories affect early perceptual stages of speech processing. In that case, acoustic cues would be interpreted in the light of the known pronunciation variants, or whether retuned categories come into play at a later post-perceptual processing stage.

The adjustment to unusual pronunciation variants of single phonemes was first demonstrated by Norris et al. (2003). They exposed Dutch listeners to a speaker who produced an ambiguous fricative between /s/ and /f/ (transcribed from here on as [$^{s}/_{f}$]). One group of participants heard this ambiguous fricative replace /s/ in s-final words, as in [mœy$^{s}/_{f}$] (“mouse”); the other group heard the same ambiguous fricative replacing /f/ in f-final words, as in [ʃirɑ$^{s}/_{f}$] (“giraffe”). Importantly, the fricative could only be interpreted as /s/ or /f/ in these stimuli since the other possible interpretation (i.e., [mœyf] and [ʃirɑs]) are nonwords in Dutch. That is, the phonetically ambiguous fricative [$^{s}/_{f}$] was presented in a lexcially unambiguous context of an existing Dutch word. Listeners could thus use lexical information to interpret the ambiguous sounds (Ganong, 1980). In a lexical decision task, which served as an exposure phase, the words with ambiguous fricatives were mostly accepted as real words. Immediately thereafter, participants had to categorize sounds along an [ɛs] to [ɛf] continuum. The results of the categorization task were influenced by the exposure condition. Participants who had heard the ambiguous fricative in f-final words gave more /f/ responses for tokens from the [ɛs] to [ɛf] continuum than participants who had heard the ambiguous fricative in s-final words. Apparently, participants had learned, guided by lexical knowledge, that the same ambiguous fricative [$^{s}/_{f}$] can be a possible implementation of either /s/ or /f/.

Further experiments on this type of perceptual learning showed that the effect is speaker specific (at least for fricatives, Eisner & McQueen, 2005; Kraljic & Samuel, 2007), but generalizes over lexical items (McQueen, Cutler, & Norris, 2006; Mitterer, Chen, & Zhou, 2011). Moreover, the effect has been shown with a variety of tasks during exposure, ranging from simply counting words (McQueen, Norris, & Cutler, 2006) to hearing a story or watching a TV show (Eisner & McQueen, 2006; Mitterer & McQueen, 2009). So it is well established that listeners flexibly retune their phoneme categories. However, there are at least two ways in which this learning might influence perception. One possibility is that the newly acquired knowledge may immediately influence the processing of incoming information in the speech stream. That is, the knowledge that the speaker produces certain sounds in an unusual fashion could be applied during the initial stage of phonetic processing, at the time when the unfolding speech signal is being processed. Alternatively, phonetic processing may not be influenced directly. Rather, the newly acquired knowledge may only be consulted after an initial, speaker-independent phonetic processing of the input. The effect of learning would then have no influence on initial phonetic processing of incoming information, but would only be integrated with the outcome of phonetic processing at a later stage.

The distinction between early versus late integration is a common one in speech perception research. Kingston and Macmillan (1995), for instance, asked whether nasalization and the first-formant frequency are perceived integrally or independently for the perception of vowel height. The research question pursued by Kingston and Macmillan was whether integration already occurs at a phonetic level or whether the dimensions are perceived independently at a phonetic level, and are integrated late at a decision level. Kingston and Macmillan used signal-detection theory to show that listeners do not distinguish between degrees of vowel nasalization and first-formant frequency but instead integrate both dimensions at a phonetic level to form one cue for vowel height. A similar question arose in the debate relating to how listeners achieve “compensation for phonological assimilation”. Phonological assimilation is a production process in which a given segment is so strongly coarticulated with its context that it “loses its identity” and takes over one property of the context segment. An example is assimilation of place of articulation of word-final nasals: an underlying /n/ in *lean bacon* /lin beɪkn/ can become an [m] in the surface form [lim beɪkn]. The underlying /n/ has then been assimilated to the labial place of articulation of the following /b/. Gaskell (2003) proposed a model of compensation for phonological assimilation in perception in which the assimilated segment (e.g., the [m] in [lim beɪkn]) is first perceived as an instance of its surface form (i.e., as /m/). Only at a later processing stage is the context taken into account, such that the [m] is treated as a possible instance of an underlying /n/. This contrasts with the proposal by Mitterer, Csépe, and Blomert (2006) who argued that the context already influences the initial perceptual processing of the assimilated segment, making the [m] “sound” like an /n/ already at an auditory level, similar to auditory backward masking (Moore, 2003).

Most prominently, the distinction between early and late integration has featured in the field of audiovisual speech perception. Proponents of gestural theories of speech perception argued that the visual and auditory information streams are integrated at an early level of speech perception (Fowler, Brown, & Mann, 2000). This contrasts with the model of Massaro (1998), in which auditory and visual sensory processing proceed independently and are only integrated at a decision stage. Massaro’s (1998) proposal—independent sensory processing in the auditory and visual domains followed by integration at a decision level—resonates with a proposal for a distinction between an initial, fast, first-pass processing and a later reevaluation in visual perception. Lamme and Roelfsema (2000) argued that there is an initial fast feedforward sweep of sensory processing that is relatively stable and automatic. Visual awareness, however, seems to depend on additional horizontal and recurrent processing, that is, processing within one brain area or re-entrant processes from later areas, respectively. As this shows, a frequent distinction is made between early first-pass sensory processing and later re-evaluation and decision processes. In a way, this distinction relates to the common title “Sensation and Perception” used for textbooks in introductory psychology.

In the current paper, we ask whether the results of lexically guided retuning of phonemes are brought to bear on