Brain & Language 157–158 (2016) 14–24

Contents lists available at ScienceDirect

Brain & Language

journal homepage: www.elsevier.com/locate/b&l

Matching heard and seen speech: An ERP study of audiovisual word recognition

Natalya Kaganovich [a,b,*], Jennifer Schumaker [a], Courtney Rowland [a]

[a] Department of Speech, Language, and Hearing Sciences, Purdue University, 715 Clinic Drive, West Lafayette, IN 47907-2038, United States
[b] Department of Psychological Sciences, Purdue University, 703 Third Street, West Lafayette, IN 47907-2038, United States

ARTICLE INFO

Article history:
Received 27 June 2015
Revised 23 March 2016
Accepted 10 April 2016
Available online 4 May 2016

ABSTRACT

Seeing articulatory gestures while listening to speech-in-noise (SIN) significantly improves speech understanding. However, the degree of this improvement varies greatly among individuals. We examined a relationship between two distinct stages of visual articulatory processing and the SIN accuracy by combining a cross-modal repetition priming task with ERP recordings. Participants first heard a word referring to a common object (e.g., pumpkin) and then decided whether the subsequently presented visual silent articulation matched the word they had just heard. Incongruent articulations elicited a significantly enhanced N400, indicative of a mismatch detection at the pre-lexical level. Congruent articulations elicited a significantly larger LPC, indexing articulatory word recognition. Only the N400 difference between incongruent and congruent trials was significantly correlated with individuals' SIN accuracy improvement in the presence of the talker's face.

1. Introduction

Seeing a talker's face considerably improves speech-in-noise (SIN) perception in both children and adults (Barutchu et al., 2010; Ross, Saint-Amour, Leavitt, Javitt, & Foxe, 2007; Sumby & Pollack, 1954; Tye-Murray, Spehar, Myerson, Sommers, & Hale, 2011), with facial speech gestures providing both redundant and complementary information about the content of the auditory signal. Indeed, recent studies show that a decrease in the SIN ratio leads to greater visual fixations on the mouth of the speaker (Yi, Wong, & Eizenman, 2013) and stronger synchronizations between the auditory and visual motion/motor brain regions (Alho et al., 2014). Importantly, however, the degree to which individuals benefit from visual speech cues varies significantly (Altieri & Hudock, 2014; Grant, Walden, & Seitz, 1998). Reasons for such variability may be many. As an example, Grant and colleagues proposed that variability in the processing of either auditory or visual modality as well as in the audiovisual integrative mechanisms may independently contribute to the degree of improvement for audiovisual as compared to auditory only speech (Grant et al., 1998).

In this study, we focused on individual variability in matching auditory words with their silent visual articulations – the skill that is at the heart of audiovisual speech perception – and asked which aspects of such matching process play a role in improved SIN perception when seeing the talker's face. Just like auditory words, visual articulations unfold over time, and their processing is incremental in nature. Viewers may detect mismatches between auditory and articulatory information in the observed facial movements on a sub-lexical level (i.e., based on syllabic and/or phonological processing) well before the completion of the entire articulatory sequence associated with a particular word. However, because many word articulations differ only in the final segments (e.g., beam vs. beet), the unequivocal decision about a match requires that the entire sequence of facial speech gestures associated with a word is completed and coincides with the articulatory word recognition. Hypothetically, either or both stages of processing facial articulatory gestures could play a role in improving SIN perception. Because facial speech gestures typically precede the onset of sound (e.g., Conrey & Pisoni, 2006; Grant, van Wassenhove, & Poeppel, 2004; McGrath & Summerfield, 1985; but see also Schwartz & Savariaux, 2014; van Wassenhove, Grant, & Poeppel, 2007), they allow listeners to make predictions about the incoming linguistic information. Higher sensitivity to correspondences between facial speech gestures and sub-lexical units may enable more accurate predictions about the auditory signal and/or a detection of a mismatch between one's prediction and the actual sound. On the other hand, within the context of discourse, the main semantic information is carried by words. It is possible, therefore, that only the recognition of the entire articulatory sequence as a word would result in greater SIN benefit.

* Corresponding author at: Department of Speech, Language, and Hearing Sciences, Purdue University, Lyles-Porter Hall, 715 Clinic Drive, West Lafayette, IN 47907-2038, United States.

E-mail address: kaganovi@purdue.edu (N. Kaganovich).

http://dx.doi.org/10.1016/j.bandl.2016.04.010
0093-934X/© 2016 Elsevier Inc. All rights reserved.

The notion that the degree of SIN improvement in the presence of the talker's face may depend on the level of linguistic analysis is supported by earlier research. For example, Grant and Seitz (Grant & Seitz, 1998) reported that their measures of audiovisual benefit for SIN during nonsense syllable and sentence perception did not correlate. In a similar vein, the study by Stevenson and colleagues (Stevenson et al., 2015) showed that healthy elderly adults benefit from visual speech cues during a SIN task as much as younger adults when presented with individual phonemes but show marked deficits when presented with individual words. While better understanding of how facial speech gestures facilitate SIN perception at different linguistic levels is needed, the above studies suggest that the mechanisms engaged at each level may be at least partially distinct.

In order to examine unique contributions of matching auditory and visual speech information at the sub-lexical and lexical level to the SIN accuracy, we combined a cross-modal repetition priming task with event-related potentials recordings (ERPs). The ERP technique's excellent temporal resolution allows one to tease apart perceptual and cognitive processes that jointly shape behavioral performance. We were, therefore, able to evaluate ERP responses associated with audiovisual matching at the sub-lexical level separately from the ERP responses associated with articulatory word recognition and correlate both measures with individuals' performance on the SIN task.

In the cross-modal repetition-priming task, participants first heard a word referring to a common object (such as a pumpkin) and then had to decide whether the subsequently presented visual silent articulation matched the word they had just heard. In half of trials, the presented articulation matched the heard words (congruent trials), and in another half it did not (incongruent trials). The important aspect of this paradigm is that in absolute terms, no trial contained a true repetition of the same physical stimulus since the first word was always presented in the auditory modality only and the second word in the visual modality only. On congruent trials, the seen articulation was expected to be perceived as a match to the auditory word and lead to the articulatory word recognition. On incongruent trials, a mismatch between the expected and the observed articulation would be detected. The ERP components associated with word repetition (including cross-modal presentations) – the N400 and the late positive complex (LPC) – have been well-studied and allow for clear predictions and interpretation of the results as described below.

The N400 ERP component is a negative waveform deflection that peaks at approximately 400 ms post-stimulus onset in young healthy adults and has a centro-parietal distribution. This component is thought to index the ease with which long-term semantic representations may be accessed during processing (for reviews, see Duncan et al., 2009; Holcomb, Anderson, & Grainger, 2005; Kutas & Federmeier, 2011; Kutas & Van Petten, 1988, 1994). However, and more germane to the topic of the current study, the N400 amplitude is also sensitive to phonological correspondences between prime and target words in priming tasks (Praamstra, Meyer, & Levelt, 1994; Praamstra & Stegeman, 1993), with greater negativity to phonological mismatches. Importantly, a study by Van Petten and colleagues demonstrated that the onset of the N400 component precedes the point at which words can be reliably recognized (Van Petten, Coulson, Rubin, Plante, & Parks, 1999), suggesting that this component is elicited as soon as enough information has been processed to determine that the incoming signal either matches or mismatches the expected one.

Based on the above properties of the N400 component, we predicted that incongruent visual articulations would elicit larger N400s compared to congruent visual articulations. Additionally, because all incongruent word pairs differed at the word onset, we expected that the N400 amplitude increase to incongruent articulations would reflect a relatively early process of detecting an expectancy violation, likely prior to the articulatory word recognition. Lastly, if sensitivity to audiovisual correspondences at the sub-lexical level plays a role in SIN perception, we expected that individuals with greater N400 differences between incongruent and congruent trials would show better improvement on the SIN task when seeing the talker's face.

The LPC ERP component belongs to a family of relatively late positive deflections in the ERP waveform that may vary in distribution and amplitude depending on the task used. Of particular relevance to our paradigm is the sensitivity of this component to word repetition (for reviews, see Friedman & Johnson, 2000; Rugg & Curran, 2007). More specifically, the LPC is larger (i.e., more positive) to repeated as compared to not repeated words (e.g., Neville, Kutas, Chesney, & Schmidt, 1986; Paller & Kutas, 1992), suggesting that it indexes some aspects of the recognition process. We hypothesized that the LPCs to congruent articulations should have larger amplitude than the LPCs to incongruent articulations, which were not expected to result in the articulatory word recognition on a regular basis. Furthermore, if recognition of the entire articulatory sequence as a specific word is important for SIN, we expected that those individuals with the largest LPC differences between congruent and incongruent articulations would show the best improvements on the SIN task when seeing the talker's face.

2. Method

2.1. Participants

Twenty-two college-age adults participated in the study for pay. They had normal hearing (tested at 500, 1000, 2000, 3000, and 4000 Hz at 20 dB SPL), normal or corrected to normal visual acuity, and normal non-verbal intelligence (Brown, Sherbenou, & Johnsen, 2010). According to the Laterality Index of the Edinburgh Handedness Questionnaire, two participants were ambidextrous, and the rest were right-handed (Cohen, 2008; Oldfield, 1971). All gave their written consent to participate in the experiment. The study was approved by the Institutional Review Board of Purdue University, and all study procedures conformed to The Code of Ethics of the World Medical Association (Declaration of Helsinki) (1964).

2.2. Stimuli and experimental design

The study consisted of two experiments. In the first (referred to henceforth as the Matching task), participants decided whether visual only articulation matched the word they had just heard. Each trial consisted of the following events (see Fig. 1). Participants first saw a color picture of a common object/person (e.g., toys, mailman, etc.). While the image was still on the screen and 1000 ms after its appearance, participants heard the object named (e.g., they heard a female speaker pronounce a word "toys" or "mailman," etc.). The image continued to stay on the screen for another 1000 ms after the offset of the sound and then disappeared. A blank screen followed for another 1000 ms. Next, a video of a female talker was presented. It consisted of a static image of the talker's face taken from the first frame of the video (1000 ms), followed by a silent articulation of a word, followed by the static image of the talker's face taken from the last frame of the video (1000 ms). In half of all trials, the talker's articulation matched the previously heard word (congruent trials; for example, participants saw the talker articulate "toys" after hearing the word "toys" earlier), while in another half, the talker's articulation clearly mismatched the previously heard word (incongruent trials;