Brain & Language 137 (2014) 86–90

Contents lists available at ScienceDirect

Brain & Language

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

Short Communication

How visual timing and form information affect speech and non-speech processing

Jeesun Kim, Chris Davis *

The MARCS Institute, University of Western Sydney, Australia

ARTICLE INFO

Article history:
Accepted 17 July 2014
Available online 3 September 2014

Keywords:
Visual speech
Auditory and visual speech processing
Visual form and timing information

ABSTRACT

Auditory speech processing is facilitated when the talker's face/head movements are seen. This effect is typically explained in terms of visual speech providing form and/or timing information. We determined the effect of both types of information on a speech/non-speech task (non-speech stimuli were spectrally rotated speech). All stimuli were presented paired with the talker's static or moving face. Two types of moving face stimuli were used: full-face versions (both spoken form and timing information available) and modified face versions (only timing information provided by peri-oral motion available). The results showed that the peri-oral timing information facilitated response time for speech and non-speech stimuli compared to a static face. An additional facilitatory effect was found for full-face versions compared to the timing condition; this effect only occurred for speech stimuli. We propose the timing effect was due to cross-modal phase resetting; the form effect to cross-modal priming.

1. Introduction

It is well established that seeing the talker's moving face (visual speech) influences the process of speech perception, e.g., speech is perceived more accurately in quiet (Davis & Kim, 2004) and in noise (Sumby & Pollack, 1954). Such visual influence has been attributed to the information available from the talker's oral regions, e.g., from mouth shapes, mouth and lip motion and some tongue positions (Summerfield, 1979) and peri-oral regions such as jaw, eyebrows and head (Davis & Kim, 2006; Munhall, Jones, Callan, Kuratate, & Vatikiotis-Bateson, 2004). The current study focused on the effect that perceiving speech-related movements has on speech processing, and was motivated by the observation that such motion provides two broad types of information, speech form (segment) and timing information (Summerfield, 1987). That is, mouth and lip movements define shapes and spaces that can combine with tongue positions to provide form information about the identity of spoken segments. In addition, such motion provides timing information about segment onset, offset and duration (Summerfield, 1979) and information about syllabic rhythmic structure from the cycle of jaw open-closure (Greenberg, Carvey, Hitchcock, & Chang, 2003; MacNeilage, 1998). We examined the extent that these two sources of speech information influence speech processing.

Understanding the influence of visual form and timing information on auditory-visual (AV) speech processing is important not only for an appreciation of each component effect, but also because explanations of AV effects have tended to emphasize the importance of either one type of information or the other. Some neurophysiological accounts see the form of visual speech as being of key importance. Take for example, the explanation that Jääskeläinen, Kauramäki, Tujunen, and Sams (2008) advanced to explain why visual speech reduced the size of the auditory N1 evoked potential to vowels. Here it was argued that seeing lip shapes from particular articulations altered the sensitivity of auditory cortical neurons responsive to frequencies in the region of the second formant. Other accounts have stressed the role that the timing plays. For example, Arnal, Morillion, Kell, and Giraud (2009) have proposed that earlier auditory evoked responses (M100) to syllables preceded by predictable visual speech was due to rhythmic information resetting the phase of oscillation of auditory cortical neurons and so by increasing their receptivity (see also Lakatos, Chen, O'Connell, Mills, & Schroeder, 2007).

Explanations of AV speech effects in behavioral studies also show a split between those that emphasize the importance of visual form and those emphasizing the importance of visual timing information. For example, explanations of the McGurk effect are typically couched with respect to the form of visual speech (McGurk & MacDonald, 1976). Likewise, it has been proposed that visual form information can reinforce or disambiguate phonemic content, especially in difficult listening environments (Hazan, Kim, & Chen, 2010) and can provide information about the spectral composition of speech (Grant & Seitz, 2000; Kim & Davis, 2004). On the other hand, it has been proposed that visual timing information also

* Corresponding author.
E-mail addresses: j.kim@uws.edu.au (J. Kim), chris.davis@uws.edu.au (C. Davis).

http://dx.doi.org/10.1016/j.bandl.2014.07.012
0093-934X/

can influence auditory speech processing. For example, a number of authors have proposed that visual speech provides cues as to when to listen (Grant & Seitz, 2000; Kim & Davis, 2004; Schwartz, Berthommier, & Savariaux, 2004).

Studies in which both form and timing information have been manipulated appear to indicate that speech form is the more important cue. For example, Paris, Kim, and Davis (2013) manipulated the form and timing information available from visual speech independently and determined how this affected the time to process a subsequently presented speech sound (to decide whether a /ba/ or /da/ was presented). Visual speech form information (showing articulation of the full face up to the point of vocalization) was presented in a random interval between 250 and 400 ms before the auditory stimulus (i.e., no reliable timing information from the visual stimulus). Visual speech timing information (showing articulation of the talker's jaw up to the point of vocalization) contained no form information about the spoken syllable. Compared to an auditory alone control, it was found that the form information significantly facilitated response times whereas the timing information did not. This result is consistent with the finding that the McGurk effect is relatively tolerant to large asynchronies between the AV speech signals, particularly when auditory speech lags (e.g., Munhall, Gribble, Sacco, & Ward, 1996).

Of course, form effects (e.g., the McGurk) are ultimately constrained by when the visual and auditory speech signals occur. Studies have shown that when presented outside a temporal window, information from visual and auditory speech does not combine to influence perception (e.g., Munhall et al., 1996). Moreover, as mentioned above, there are studies that make it clear that the timing of visual speech information is crucial to its influencing auditory speech processing. For example, Kim and Davis (2004) have reported that in a speech detection in noise task the boost in accuracy due to presenting visual speech was eliminated by misaligning the AV signals by even a relatively small margin (40 ms).

One way of interpreting the divergent results regarding the relative importance of visual speech form and timing information is to assume that the nature of the task used to measure speech processing modulates the degree of precision required from these different information types. Identification and detection tasks differ in terms of the properties of stimulus they employ and also at the level of processing used to drive responses. Identification tasks (identifying words or speech segments) use largely intact speech signals and responses are determined at a relatively late stage of processing. These tasks appear to be more sensitive to visual form information. Detection tasks typically present a severely degraded speech signal where participants are required to identify when (or if) a stimulus has occurred. It has been suggested that this type of task taps early stages of stimulus processing (Grant & Seitz, 2000). Detection tasks tend to show that the synchrony (timing) of visual to auditory speech is important. Given the potential importance of task, the current experiment used a task that combined properties of detection (detecting some relatively basic properties) and identification (recognizing a class of object) to jointly examine the contribution of visual form and timing information to speech processing.

In relation to a visual timing effect, our interest was to determine whether seeing peri-oral speech-related motion would facilate processing speech, non-speech or both. Previous studies have found that visual speech can facilitate target speech detection (Kim & Davis, 2003, 2004) or that the simultaneous presentation of a visual cue (a light) can improve the detectability of a target sound (Lovelace, Stein, & Wallace, 2003) when the target is heavily masked or presented at threshold. However, it is not clear that such a timing effect will occur in current setup where a clear target signal was not heavily masked and the task did not involve auditory detection (see below). Few experiments have been conducted on whether visual timing cues can assist auditory identification and the results have been mixed. For instance, Schwartz et al. (2004) showed that the presentation of a visual stimulus that provided speech timing but no speech form information (a rectangle that increased and decreased in height according to measures of mouth articulation) did not improve speech intelligibility. On the other hand, Best, Ozmeral, and Shinn-Cunningham (2007) showed that a cue (switching on LEDs) that indicated the time that a target would occur in a complex acoustic mixture improved identification accuracy and that this occurred for both speech and non-speech signals. Best et al. suggest their results may have been due to phasic alerting where the visual stimulus facilitated auditory processing by directing attention to the appropriate point in time.

The aim with respect to visual speech form was to determine if it primes the processing of corresponding auditory speech. This was examined by contrasting stimuli where the auditory and visual speech matched with those where they did not. This manipulation was based on demonstrations that there is a functional correspondence between lip and mouth movements and particular speech spectral properties (Berthommier, 2004; Girin, Schwartz, & Feng, 2001) and that seeing visual speech significantly up-regulates the activity of auditory cortex compared to auditory speech alone (Okada, Venezia, Matchin, Saberi, & Hickok, 2013). Combining these two observations leads to the prediction that visual speech form will facilitate decisions based on the processing of its auditory counterpart.

In what follows we outline the factors that were considered in designing the task to be used and the method for presenting timing and form information. First, the task required that responses were based on detecting speech (thus potentially sensitive to the timing of visual speech at an early stage of speech processing). In this regard, a task was selected that required a simple binary speeded response based on whether the presented stimulus sounded like speech or not. Second, the speech and non-speech stimuli differed in their spectral distribution but not in their temporal structure, i.e., the non-speech stimuli were created from the speech ones by spectral rotation (see Method). Thus discriminating the speech from the non-speech stimuli required that participants identify the spectral signature of speech (it is this signature that should correspond to the information presented in the visual speech form). Third, the speech stimuli consisted of nonwords in order to minimize the influence of lexical processing (i.e., high level information processing). In addition, all stimuli were presented in a moderate level of noise (−5 dB) so that there was uncertainty when the signal started but the signal itself was relatively intact. The video began with the noise playing (before the target sound was presented) so that participants had time to prepare their response (rather than potentially reacting to the sudden onset of the sound).

In the experiment, visual speech timing information was presented by showing the talker's peri-oral movements with the mouth area obscured by an overlaid opaque circular patch (see

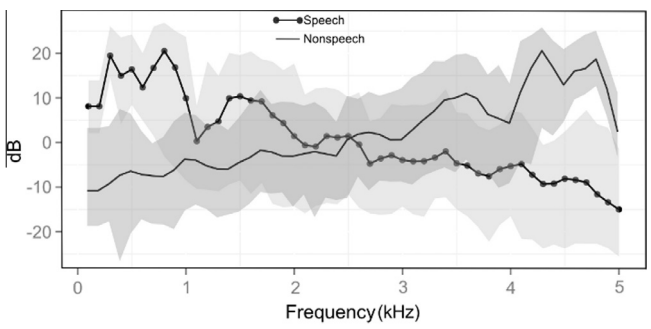

Fig. 1. The long-term average spectrum (LTAS) of the nonword speech and non-speech stimuli. The curves represent the mean LTAS for the 45 stimuli in each condition, the shaded grey ribbons indicate the range.