Please cite this article in press as: Paris T et al. Visual form predictions facilitate auditory processing at the N1. Neuroscience (2016), http://dx.doi.org/10.1016/j.neuroscience.2016.09.023

Neuroscience xxx (2016) xxx–xxx

VISUAL FORM PREDICTIONS FACILITATE AUDITORY PROCESSING AT THE N1

TIM PARIS,* JEESUN KIM AND CHRIS DAVIS

The MARCS Institute, University of Western Sydney, Sydney, Australia

Abstract—Auditory-visual (AV) events often involve a leading visual cue (e.g. auditory-visual speech) that allows the perceiver to generate predictions about the upcoming auditory event. Electrophysiological evidence suggests that when an auditory event is predicted, processing is sped up, i.e., the N1 component of the ERP occurs earlier (N1 facilitation). However, it is not clear (1) whether N1 facilitation is based specifically on predictive rather than multisensory integration and (2) which particular properties of the visual cue it is based on. The current experiment used artificial AV stimuli in which visual cues predicted but did not co-occur with auditory cues. Visual form cues (high and low salience) and the auditory-visual pairing were manipulated so that auditory predictions could be based on form and timing or on timing only. The results showed that N1 facilitation occurred only for combined form and temporal predictions. These results suggest that faster auditory processing (as indicated by N1 facilitation) is based on predictive processing generated by a visual cue that clearly predicts both what and when the auditory stimulus will occur. © 2016 Published by Elsevier Ltd on behalf of IBRO.

Key words: audiovisual, prediction, N1 latency, EEG.

INTRODUCTION

In many ecological settings, multisensory signals that indicate the presence and identity of an object/event occur at similar times and provide redundant and sometimes complementary information. Research investigating the sensory, perceptual and cognitive processing of these multisensory cues has largely focused on the combination of information, so what has often been overlooked is the role that the temporal order of these cues has on processing. That is, it is commonplace with multisensory signals for a cue from one modality to precede the other, for example, in the case of auditory-visual (AV) speech, the movements of the lips and jaw often begin before the auditory signal is produced. Recent research indicates that prior visual information can be used to generate a prediction about the upcoming sound such as what (Kim and Davis, in press; van Wassenhove et al., 2005), when (Vroomen and Stekelenburg, 2010) and where (Stekelenburg and Vroomen, 2012) it will occur, and this information results in changes to subsequent auditory processing.

In this regard, one particularly interesting suggestion has been that predictions derived from the visual modality can speed up auditory processing (van Wassenhove et al., 2005; Paris et al., 2013). This suggestion was based on the finding that the N100 (N1) ERP in response to auditory speech occurs earlier when preceded by visual speech (Arnal et al., 2009; van Wassenhove et al., 2005). Although it was known that auditory stimulus features (e.g., intensity, Jacobson et al., 1992) and auditory expectations (Budd and Michie, 1994) influence N1 latency, what was intriguing about this 'N1 facilitation effect' was that it demonstrated that predictions generated in one modality (visual) could facilitate processing in another (auditory). Further, it was suggested that the visually induced shift in auditory N1 latency is caused by AV interactions in sensory cortices (Besle et al., 2004; van Wassenhove et al., 2005; Arnal et al., 2009).

The AV interactions that give rise to the N1 facilitation effect have been suggested to reflect predictive processing and not that of multisensory integration per se. The argument for this is that the onset latency of the N1 is affected by the characteristics of the visual signal that begins prior to the acoustic event and not by the relationship between auditory and visual cues that is likely determined during and/or after the auditory event. For example, Arnal and colleagues (2009) have demonstrated that more salient visual speech (i.e., well-marked, distinct movements of the lips and jaw, associated with the articulation of a particular sound, e.g., 'pa') resulted in an earlier N1 latency compared to less salient visual speech (e.g., 'ga') and the validity of the prediction did not influence the amount of N1 facilitation. That is, an invalid prediction (such as in the lip movements of 'pa' paired with the sound of 'ga') resulted in the same latency facilitation as a valid prediction (visual 'pa' paired with auditory 'pa').

Although these results have been interpreted as indicating that N1 facilitation is due to visually based prediction, one feature of these experiments potentially undermines this interpretation. Typically, experiments have used AV signals that overlap in time,

*Corresponding author. Address: The MARCS Institute, University of Western Sydney, Bullecourt Avenue, Milperra 2214, NSW, Australia. E-mail address: t.paris@uws.edu.au (T. Paris).
Abbreviations: AV, auditory-visual; RT, response times; VO, visual only.

http://dx.doi.org/10.1016/j.neuroscience.2016.09.023

i.e., participants saw visual speech that occurred both prior to and during auditory speech presentation. Since the AV signals overlapped, it cannot be unequivocally concluded that N1 facilitation occurs purely due to visually based predictions.

Further, even if N1 facilitation is due to prediction based on the leading signal, other interpretative issues remain. First, it is unclear whether prediction per se is sufficient to produce N1 facilitation or whether the pairing of the AV signals first needs to be over-learned. That is, N1 facilitation has almost exclusively been demonstrated with AV stimuli that have been extensively learnt, such as AV speech stimuli or with other ecological stimuli such as clapping hands (Stekelenburg and Vroomen, 2007). It therefore is not clear whether N1 facilitation requires pre-established AV pairings or whether it would occur whenever the presentation of one stimulus reliably predicts the occurrence of another. The former proposal is consistent with the suggestion that N1 facilitation occurs for speech stimuli because lip movements and speech sounds have a well-learned and tight mapping that allow predictions to be established (van Wassenhove et al., 2005).

The other interpretive issue is about which features of the visual prediction are important in triggering N1 facilitation effects. Previous research has demonstrated how either form or temporal predictions can influence early neural responses using non-speech stimuli. Firstly in terms of form predictions, a number of studies used simple AV associations to investigate the effect of prior visual cues on auditory evoked responses (Widmann et al., 2004; Laine et al., 2007; Lindström et al., 2012). For example, in the study by Widmann and colleagues, participants were presented with horizontal bars of varying heights that were associated with the pitch of subsequent tones. The results of this study showed that valid visual form cues reduced the amplitude of the N1 relative to invalid ones, yet importantly there was no evidence of a latency shift in N1 (see also Laine et al., 2007; Lindström et al., 2012).

In terms of temporal prediction, a study by Vroomen and Stekelenburg (2010) presented participants with moving disks that collided with a central rectangle at which time a sound was produced. When these disks reliably preceded the sound (i.e., the disks provided a prediction as to when the sound would occur) a small N1 facilitation effect occurred. However, this effect was not robust as a second experiment using the same methods did not replicate this finding. Further, it is unclear whether the observed N1 facilitation effect in the study was due to temporal prediction alone as only a single tone was used in this experiment, thus also rendering the form predictable. Given this, whether temporal predictions alone can induce N1 facilitation is yet to be confirmed. Taken together, the above studies suggest that N1 facilitation does not occur for form-only predictions and it is unclear whether facilitation occurs for temporal-only predictions. In light of this it should be noted that ecological stimuli that do show facilitation effects contain both form and temporal cues (Paris et al., 2013).

The current study will address the above issues using non-ecological (artificial) AV stimuli (expanding shapes and tones, see below) that allow complete control of stimulus parameters. We designed visual stimuli that contained key features of AV speech, i.e., the form of the stimuli evolved over time to enable the prediction of the onset and type of the upcoming sound, but unlike speech the visual and auditory components did not overlap. In addition, we also controlled other features such as the time-course and salience of the visual prediction. In this way we could test three main questions: First, whether dynamic visual form predictions that occur prior to the sound would facilitate N1 latency; second, whether this effect would be moderated by stimulus form salience; and third, whether timing information only would facilitate responses.

EXPERIMENTAL PROCEDURES

Participants

Seventeen female participants from the University of Western Sydney took part in the experiment. Their age ranged from 17 to 40 with a mean age of 25 years. All participants reported having normal or corrected-to-normal vision and hearing and were right-handed. The study was conducted with approval of the ethics committee of the University of Western Sydney.

Experimental design and stimuli

The functional properties of the audiovisual stimuli were designed to be similar to those of ecological stimuli (i.e., a brief dynamic visual stimulus that changed to signal the onset of a specific sound). To satisfy these requirements, three types of videos were created that each consisted of a different expanding visual shape that preceded an auditory event. Each type began with a small fixation circle (shown centered in a square), which lasted for a variable window of 300 to 1100 ms. The shape then expanded into one of three possible shapes (sharp, round and rounded diamond, see Fig. 1B) occurred for 500 ms, whereupon the shape made contact with the edge of the square and disappeared as a high (1000 Hz) or low tone (333 Hz) played (for a duration of 100 ms and a rise/fall-time of 10 ms). An example of the time-course of an AV trial is shown in Fig. 1.

In the study there were two unimodal (AO and VO) and three AV conditions. The AO condition consisted of a display of a static fixation circle followed by one of the two tones and the VO condition consisted of a video with no sound. The AO condition acted as the 'unpredicted' stimuli as no visual cue preceded the sound. The AV condition consisted of two types of form and temporal prediction cues: (AVvalid) in which visual form and timing cues validly predicted a tone or (AVinvalid) where the form and timing cues provided an invalid prediction. The third AV condition consisted of a visual cue that only provided reliable temporal information (AVtemp). In order to create predictions from the visual stimuli, we manipulated the proportion of