Brain and Cognition 75 (2011) 60–66

Contents lists available at ScienceDirect

Brain and Cognition

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

# Effects of audio–visual integration on the detection of masked speech and non-speech sounds

Ranmalee Eramudugolla [a,*], Rachel Henderson [b], Jason B. Mattingley [a,b]

[a] Queensland Brain Institute, The University of Queensland, St Lucia, Queensland 4072, Australia
[b] School of Psychology, The University of Queensland, St Lucia, Queensland 4072, Australia

ARTICLE INFO

*Article history:*
Accepted 24 September 2010
Available online 9 November 2010

*Keywords:*
Multisensory integration
Audio–visual
Speech
Auditory perception
Visual perception
Lip-reading

ABSTRACT

Integration of simultaneous auditory and visual information about an event can enhance our ability to detect that event. This is particularly evident in the perception of speech, where the articulatory gestures of the speaker's lips and face can significantly improve the listener's detection and identification of the message, especially when that message is presented in a noisy background. Speech is a particularly important example of multisensory integration because of its behavioural relevance to humans and also because brain regions have been identified that appear to be specifically tuned for auditory speech and lip gestures. Previous research has suggested that speech stimuli may have an advantage over other types of auditory stimuli in terms of audio–visual integration. Here, we used a modified adaptive psychophysical staircase approach to compare the influence of congruent visual stimuli (brief movie clips) on the detection of noise-masked auditory speech and non-speech stimuli. We found that congruent visual stimuli significantly improved detection of an auditory stimulus relative to incongruent visual stimuli. This effect, however, was equally apparent for speech and non-speech stimuli. The findings suggest that speech stimuli are not specifically advantaged by audio–visual integration for detection at threshold when compared with other naturalistic sounds.



## 1. Introduction

Integration of inputs from different senses which correspond with a single event or object is important for several reasons, one of which is resolving ambiguous or degraded signals (Stein & Stanford, 2008b). Critical cues to whether stimuli from separate sensory modalities apply to a single event include spatial and temporal coincidence of the stimuli, as well as semantic congruence. Multisensory integration is apparent when input from one sensory modality enhances the perception of stimuli in another modality. For example, observers' threshold for detecting a speech stream embedded in noise is *lowered* if they are also watching a video of the speaker's articulatory movements, relative to when no video is presented. In contrast, the detection threshold is *increased* if the speech stream is paired with a video that does not correspond to the speech stream, or that matches the speech stream but is temporally out of phase (Bernstein, Auer, & Takayanagi, 2004; Grant & Seitz, 2000; Kim & Davis, 2004). Recognition of speech in noise is similarly improved by presenting matched, synchronous visual stimuli relative to presenting the auditory stimuli alone (Ross, Saint-Amour, Leavitt, Javitt, & Foxe, 2007; Sumby & Pollack, 1954). Multisensory enhancement of auditory or visual perception can also be achieved for simpler, non-speech stimuli by synchronously presenting tones or noise bursts with flashes of light or shape stimuli (e.g., Giard & Peronnet, 1999; Miller, 1991).

It has been claimed, however, that the integration of visual articulatory gestures with auditory speech represents a special case of multisensory integration, and that additional speech-specific mechanisms, and brain networks, may act to enhance the integration of speech stimuli over other categories of sound (Calvert et al., 1997; Campanella & Belin, 2007; Jones & Jarick, 2006; Saldana & Rosenblum, 1993). The level of processing at which this speech-specific effect occurs is not clear, and most studies that have compared speech and non-speech stimuli have focussed on audio–visual integration during stimulus recognition and identification. Neurophysiological and neuroimaging data suggest that visual and auditory speech may be integrated at very early levels of processing (Bernstein et al., 2004; Calvert & Campbell, 2003; Calvert et al., 1997; Ghazanfar, Maier, Hoffman, & Logothetis, 2005; Pekkola et al., 2005). If specialised integration effects occur early in the processing of auditory and visual stimuli, then such multisensory effects should be observed for the detection of speech signals and not just their identification. Here, we compare the effects of audio–visual integration of speech and non-speech stimuli on the detection of noise-masked auditory stimuli.

* Corresponding author.
*E-mail address:* r.eramudugolla@uq.edu.au (R. Eramudugolla).

0278-2626/$ - see front matter © 2010 Elsevier Inc. All rights reserved.
doi:10.1016/j.bandc.2010.09.005

Visual influences on auditory speech processing can be observed within early auditory sensory regions. However, it is unclear whether this reflects audio–visual integration at very early stages of sensory processing, or feedback modulation of early areas following integration at later, lexical/semantic processing levels. Several fMRI studies have reported that visual perception of articulatory gestures in the absence of auditory speech can activate auditory regions such as the superior temporal sulcus (Bernstein et al., 2002; Calvert & Campbell, 2003; Calvert et al., 1997) and even primary auditory cortex (Pekkola et al., 2005 although see Bernstein et al., 2002). These responses to visual speech were significantly stronger than for other types of non-speech dynamic visual stimuli incorporating faces (Pekkola et al., 2005). Indeed, there is increasing evidence for converging visual and auditory inputs within early sensory processing areas (Macaluso & Driver, 2005; Stein & Stanford, 2008a) and the results of electrophysiological studies suggest audio–visual interaction can occur even at very early stages of sensory processing (40–50 ms post-stimulus) – at least for simple auditory tones and visual disc stimuli (Giard & Peronnet, 1999; Molholm et al., 2002).

Electrophysiological studies of audio–visual integration of speech show that visual speech stimuli significantly modulate early responses to auditory speech stimuli. Specifically, the early auditory N1 response to speech syllables have a smaller amplitude (Besle, Fort, Delpuech, & Giard, 2004; Stekelenberg & Vroomen, 2007; van Wassenhove, Grant, & Poeppel, 2005) and an earlier latency (van Wassenhove et al., 2005) when presented under audio–visual conditions compared to unimodal auditory conditions. This decrease in the auditory N1, which typically peaks between 100 and 150 ms post-stimulus, is thought to reflect multisensory integration during the initial, feed-forward processing of sensory information (Besle & Giard, 2009). Critically, studies that have directly compared speech with non-speech stimuli suggest that this audio–visual effect on early auditory ERPs is not specific to speech because the decrease in auditory N1 was observed for both speech and non-speech stimuli (Klucharev, Mottonen, & Sams, 2008; Stekelenberg & Vroomen, 2007). These studies used suprathreshold stimuli with visual movements that preceded the auditory stimulus by 160–320 ms, in both the speech and non-speech stimuli. In fact, the audio–visual modulation of the auditory N1 was found to be related to the presence of anticipatory visual motion rather than congruence of the auditory and visual stimulus pairings. Stekelenberg and Vroomen (2007) suggested that under natural circumstances, early audio–visual interactions may involve visual input cueing the auditory stimulus – however, a purely attentional effect on auditory N1 was ruled out because directing attention to the auditory modality typically increases, not decreases, early auditory ERPs. Although the above studies observed multisensory effects on ERPs, effects on behaviour could not be ascertained because participants were not engaged in a specific detection or identification task on the presented stimuli. In order to understand whether multisensory integration of speech and non-speech differ at early levels of processing, it is necessary to compare the effects at the behavioural level as well.

Behavioural investigations of speech-specific advantages in multisensory integration have largely focussed on stimulus recognition and identification. Such studies generally report that audio–visual integration benefits auditory speech perception to a greater extent than other types of sounds. The temporal window within which audio–visual speech can be integrated is larger than that for non-speech stimuli, and this implies that integration occurs more readily for speech than other types of stimuli. The onset asynchrony between auditory and visual speech stimuli can be as large as 80–160 ms (audition lagging vision) before normal observers can detect the asynchrony, and multisensory enhancement of speech intelligibility is sustained with asynchronies up to about 80 ms (see Summerfield, 1992). In contrast, observers appear to be able to detect smaller audio–visual asynchronies (60–70 ms) when tested with non-speech, click and flash stimuli (Zampini, Shore, & Spence, 2003). Vatakis and Spence (2007) reported that when brief auditory and visual speech signals are presented with varying onset asynchrony, observers are poorer at judging their temporal order when the two stimuli match (i.e., pertain to the same syllable) than when the two stimuli do not match. In contrast, in a subsequent study (Vatakis & Spence, 2008) they found no difference in the judgment of temporal order for matched vs unmatched naturalistic non-speech stimuli (e.g., notes played on a musical instrument, or a person hammering). They concluded that observers are better able to integrate audio–visual speech, and that this might reflect a combination of 'top-down' factors such as greater familiarity and salience of speech stimuli, as well as more automatic processes such as temporal integration. Examining audio–visual interactions close to the threshold of stimulus detection may minimise the effects of such top-down factors.

Very few studies have examined the role of audio–visual speech integration on auditory detection at threshold. Bernstein et al. (2004) employed an adaptive staircase paradigm to examine mechanisms of audio–visual speech integration in masked speech detection. In this study, participants were required to judge, with a forced-choice response, which of two noise-masked stimulus intervals contained the spoken consonant–vowel combination 'ba'. Detection of the speech signal was examined under a unimodal auditory condition and three audio–visual conditions where the speech was paired with either a synchronised video of matching lip movements, or a video of an oval shape animated to correlate with the amplitude of the signal, or a static rectangle. The threshold signal to noise ratio (SNR) for detecting the speech signal was significantly enhanced by presenting a visual stimulus relative to the unimodal condition, and was further enhanced by presenting the lip-movement video compared with the amplitude-correlated animated shape or the static rectangle. However, excluding preliminary mouth gestures from the video of lip movements abolished this audio–visual speech advantage, indicating that amplitude correlation alone does not support audio–visual speech integration as has been previously suggested (Grant & Seitz, 2000). Although the dynamic shape stimulus controlled for audio–visual amplitude correlation, it was visually far less complex than the lip-movement video. Also, Bernstein et al. (2004) did not compare the audio–visual advantage for detecting masked speech with that for detecting other types of masked naturalistic sounds. In order to ascertain whether there is a speech-specific advantage in audio–visual interaction for stimulus detection, it is important that the integration of speech is compared with non-speech stimuli that are similarly dynamic, naturalistic, and comprised of meaningfully related visual as well as auditory components.

In the present study we used an adaptive staircase approach to examine the audio–visual detection advantage for speech and non-speech stimuli. We employed naturalistic speech and non-speech audio–visual stimuli. On a given trial of the adaptive staircase, the auditory component of a stimulus was presented in one of two noise-masked intervals (Fig. 1). The participants were required to judge, with a forced-choice response, which of the two noise intervals contained the auditory stimulus/signal. We used a three up, one down adaptive staircase where the level of the signal was decreased relative to the noise by one step (e.g., 1 dB) following three consecutive correct responses, and increased by one step following a single incorrect response (Fig. 2). This method converges on a signal/noise ratio (SNR) that produces a 79.4% rate of detection for the participant. A video stimulus was also presented on each trial, synchronised with the onset of the noise-only and noise + signal intervals. audio–visual congruency was manipulated such that on half the trials, the video matched the auditory