Investigating the effect of visual phonetic cues on the auditory N1 & P2

Daniel Hochstrasser
(BA-Psych)

An empirical thesis in accordance with the requirements of the degree
MASTERS OF RESEARCH

WESTERN SYDNEY
UNIVERSITY

The MARCS Institute for Brain,
Behaviour and Development

Acknowledgements

I would like to express my sincere gratitude to Prof. Jeesun Kim and Prof. Chris Davis for their guidance and expertise during the Masters of Research. I would also like to thank my colleagues in the Multisensory Communication team of the MARCS Institute for their patience. Finally, I would like to thank my family for their unwavering support through each new endeavour.

Abstract

Studies have shown that the N1 and P2 auditory event-related potentials (ERPs) that occur to a speech sound when the talker can be seen (i.e., Auditory-Visual speech), occur earlier and are reduced in amplitude compared to when the talker cannot be seen (auditory-only speech). An explanation for why seeing the talker changes the brain's response to sound is that visual speech provides information about the upcoming auditory speech event. This information reduces uncertainty about *when* the sound will occur and about *what* the event will be (resulting in a smaller N1 and P2, which are markers associated with auditory processing). It has yet to be determined whether form information alone can influence the amplitude or timing of either the N1 or P2. We tested this by conducting two separate EEG experiments. In Experiment 1, we compared the N1 and P2 peaks of the ERPs to auditory speech when preceded by a visual speech cue (Audio-visual Speech) or by a static neutral face. In Experiment 2, we compared contrasting N1/P2 peaks of the ERPs to auditory speech preceded by print cues presenting reliable information about their content (written "ba" or "da" shown before these spoken syllables), or to control cues (meaningless printed symbols). The results of Experiment 1 confirmed that the presentation of visual speech produced the expected effect of amplitude suppression of the N1 but the opposite effect occurred for latency facilitation (Auditory-only speech faster than Audio-visual speech). For Experiment 2, no difference in the amplitude or timing of the N1 or P2 ERPs to the reliable print versus the control cues was found. The unexpected slower latency response of the N1 to AV speech stimuli found in Experiment 1, may be accounted for by attentional differences induced by the experimental design. The null effect of print cues in Experiment 2 indicate the importance of the temporal relationship between visual and auditory events.

Author's Declaration

The work presented in this thesis is, to the best of my knowledge and belief, original except as acknowledged in the text. I hereby declare that I have not submitted this material, either in full or in part, for a degree at this or any other institution.

[redacted]……………………......

(Signature)

Table of Contents

List of Figures and Tables

Figures………………………………………………………………………………………

Tables………………………………………………………………………………………...

Introduction

This thesis explores the neural correlates of Auditory-Visual (AV) speech processing. Specifically, it examines how the onset and size of the auditory N1 and P2 event related potentials (ERPs) are affected by the different types of predictive cues. Before describing the details of the study, it is useful to situate the project within the general context of multisensory processing and the theory of predictive coding.

Our perceived reality is of a singular, connected, and coherent external world. However, incoming sensory input is separate and consists of properties that differ for each modality. A challenge of modern psychology and neuroscience is to describe the mechanisms that combine these distinctly separate and ambiguous sensations into a unified whole. An illustrative example of these mechanisms at play is the naturally occurring multisensory process of speech perception. Visual speech, the visible articulation of the lips, jaw, and tongue during speech production, directly affects auditory speech processing (Summerfield, 1992). This is usually displayed as marked improvements in accuracy when identifying auditory speech sounds (Sumby & Pollack, 1954), however, visual speech can also alter our perception of the auditory signal. For instance, when speech syllables are presented with a different visual syllable (e.g., /ga/) to that of the auditory syllable (e.g., /ba/) the two are typically merged into a novel AV percept (/da/) (MacDonald & McGurk, 1978). This phenomenon, the McGurk effect, demonstrates how our perceptual system, when attempting to process two discrepant signals, will reduce ambiguity by merging characteristics from both modalities to generate a new percept. The neural process of receiving unimodal sensory inputs and forming them into a coherent, combined and new mental representation is called, Multisensory Integration (Stein et al., 2010). A classical explanation for multisensory integration is that our unimodal sensory systems are processed separately and then combined at a later stage within specific regions of the brain devoted to the integration process (Meredith & Stein, 1986).

Contrary to this early conception, current research using electrophysiological measures (Electroencephalography, EEG or Magnetoencephalography, MEG) has revealed that multisensory processing can occur at relatively earlier stages of processing, e.g., 40-250ms after stimulus presentation (Giard & Peronnet, 1999). An alternative proposal suggests that the brain processes the environment utilising a form of Bayesian probability to generate models of the external world (Friston, 2005; Rao & Ballard, 1999). Under this theory sensory input is organised internally into relationships of cause and effect. Top-down predictions of upcoming sensory input are fed forward to lower sensory mechanisms. These predictions reduce the sensory load required to process expected signals. This theory, predictive coding, has been expanded to explain various multisensory integration effects that occur at early neural stages. Once again turning to the example of AV speech perception, past research has outlined three main predictive cues inherent in the AV speech relationship. Summerfield (1987) theorised that visual speech cues the perceiver to when the sound will occur (timing), where the sound will be (space), and what sound will be produced (form).

Researchers studying the effects of multisensory integration using EEG have shown that visual speech presented in conjunction with auditory speech stimuli reduces the magnitude of event related potentials (ERP) to the speech sounds and speeds-up when such potentials occur compared to auditory-only (AO) speech stimuli (Baart, 2016). However, there is still debate about which predictive characteristic (timing, space, or form) of the visual speech signal is influencing these neural effects. One argument is that these effects at 40 – 150 ms are too early to reflect phonetic processing and it is the temporal and spatial characteristics of visual speech which are influencing any such AV speech effect (Klucharev, Möttönen, & Sams, 2003; Stekelenburg & Vroomen, 2007). Another theory suggests that the phonetic (form) cues provided by visual speech act to reduce the amount of phonetic processing required by priming the neuronal populations associated with each phoneme before the auditory signal is processed (Besle, Bertrand, & Giard, 2009).

A reason for why it has been difficult to determine the precise effects of the various visual speech cues is due to the nature of these cues themselves. The temporal, spatial and form characteristics of visual speech are intertwined in the visual speech signal (Summerfield, 1992) making it difficult to manipulate and examine specific AV relationships. In the current study, we have adopted a different approach to the problem. Rather than try to tease apart the individual properties of visual speech cues, we have developed an entirely new AV relationship with a different visual cue: orthography or print stimuli to serve as a function equivalent for some properties of visual speech stimuli. The benefit of using printed stimuli is that there is an already established relationship between each symbol and its phonemes within a language. To test whether auditory speech processing can be influenced by phonetic information alone we will use print stimuli to provide predictive cues to the content of an auditory speech signal. This thesis consists of two EEG studies in which we aim to determine whether visual form cues alone can influence auditory speech perception at a relatively early stage of processing (40-300 *ms*). The following chapters provide an overview of the predictive coding hypothesis and the neurophysiological effects of AV integration. Afterwards, two experimental studies are described and their results explained within the context of Predictive Coding.

The Predictive Coding Hypothesis

The computational theory of neurobiological processes, Predictive Coding, provides a description for how lower-order sensory processes may interact with higher-order cognitive mechanisms. The central thesis of Predictive Coding is that the brain generates top-down predictions about upcoming sensory inputs through an internal model that archives statistically predictable causal relationships of external events (Friston, 2005; Rao & Ballard, 1999). Under this computational model, bottom-up and top-down neural pathways operate in both a hierarchical and a bi-directional order. The outline of this model is as follows: information of external events is received through the sensory system and fed bottom-up to our higher order processes. As we perceive this external data, an internal model is generated which evaluates the relationship between separate events. Events that have a high degree of co-occurrence and thus, a high predictability, are utilised to form expectations about upcoming signals. Predictive information about the expected signal is fed forward from the top-down internal model to our lower sensory mechanisms. These predictions are then used to suppress the incoming sensory information that is no longer required for perception. Any errors or violations in these predictions are fed back into the internal model. The model is subsequently updated enabling it to formulate a higher degree of accurate predictions in the future (for a detailed review see Clark, 2013). One hypothesised benefit of this computational model is a marked reduction in the cognitive load required to process a noisy, complex, and ambiguous external environment (Friston, 2005).

When applying the perceptual model of Predictive Coding to mechanisms of multisensory perception the first step is to determine a multimodal relationship that signals predictable events between the two modalities. Researchers have investigated multimodal predictive relationships within AV Speech (Paris, Kim, & Davis, 2016b; Stekelenburg & Vroomen, 2015). One reason for this, is that auditory speech is almost always paired with visual speech. A requirement of an accurate internal predictive model is that the external co-occurring relationship has a high degree of consistency. Second, the naturally occurring articulatory movements provided by visual speech typically occur before auditory speech (Chandrasekaran, Trubanova, Stillittano, Caplier, & Ghazanfar, 2009). Although, it was recently determined that the most useful informational aspects of the visual speech signal overlapped with the onset of the acoustic speech signal (Venezia, Thurman, Matchin, George, & Hickok, 2016). Nevertheless, this visual lead allows the system to make reliable predictions about the onset time of auditory information based on these visual cues (Temporal cueing) (Summerfield, 1992). The visual movements of the speech articulators also provide an indication from where sounds will occur (Spatial cueing), allowing the perceiver to attend to the speaker's location. Finally, the lip/mouth configurations created with speech give an early guide to the phonemes that will be uttered (Form Cueing). One way researchers have attempted to investigate the effects of predictive cues in AV speech is by measuring neurophysiological markers of auditory processing.

Measuring Neurophysiological effects of AV integration

There are two common measures of neural processing used in neurocognitive research. The first, functional magnetic resonance imaging (fMRI), allows us to observe neural functions by measuring an influx of oxygenated blood to a cortical area that occurs neural/cognitive processing. The underlying logic of this method is that by utilising cognitive resources the brain requires oxygenated blood to be resupplied into exhausted areas. This allows to the identification of neuronal populations within the brain that function during cognitive processes related to what a person does in an experimental task. fMRI studies have determined that AV speech integration occurs within the Superior temporal gyrus/sulcus, the Left middle temporal gyrus, and the auditory cortex (Callan et al., 2004). Despite fMRI's utility to measure the spatial position of cognitive processes, this method has a poor temporal resolution. The reliance on hemodynamic response times limits the period of measurable (a minimum of 3s – 6s after stimulus event) neuronal activity (Glover, 2011; Sejnowski, Churchland, & Movshon, 2014). Thus, fMRI does not provide the requisite temporal precision for measuring prediction effects, that can occur 40-300ms after stimulus presentation. To understand the time course of processes in neuronal populations when responding to a stimulus, researchers opt to use electrophysiological measures of brain function.

Electrophysiological methods (EEG or MEG) are useful in determining when and how neuronal processes are produced (Murali & Kulish, 2007). EEG measures the electrical activity that is propagated across large populations of neurons via post-synaptic connections and action potentials. A common method for measuring the effects of AV integration is by generating Event-Related Potentials (ERP). To generate an ERP, stimuli are presented multiple times to elicit numerous responses. Responses are averaged together to generate an ERP. The resulting waveform is considered a representation of the electrical fluctuations that occur during the make-up of an ERP contains various common components in a waveform depending on the characteristics of the stimuli. Responses to auditory stimuli influenced by visual information generate effects that manifest in 2 common components: The N1 and P2 peaks of an auditory ERP (Baart, 2016; Klucharev et al., 2003; Näätänen & Picton, 1987; Picton, Hillyard, Krausz, & Galambos, 1974).

Characteristics of The N1 and P2 peaks

The N1 (or N100) and the P2 (or P200) are two peaks of an ERP that are consistently evoked from the presentation of auditory stimuli (Picton et al., 1974). The N1 is the first negative deflection within an auditory ERP that is elicited around 50ms to 150ms after stimulus presentation (Näätänen & Picton, 1987). The N1 is not a unitary phenomenon, and as such it has more than a single generator - The first component is a frontocentral negativity (N1b) occurs at 80 to 100 ms and is probably generated by bilateral vertically oriented dipoles near the primary auditory cortices; The second component is the so called biphasic T-complex, (Wolpaw and Penry, 1975). It has a positive wave at 100 ms and a negative wave at

150 ms. This complex probably originates bilaterally in the auditory association cortex in the superior temporal gyrus. The third component tends to be a nonspecific response generating a vertex-negative wave at about 100 ms (possibly located in the frontal motor and/or premotor cortex) (Wolpaw and Penry, 1975).The P2 or P200 is a positive deflection in an ERP that occurs between 150 – 250ms after the stimulus onset (Picton et al., 1974). The N1/P2 complex are referred to as such based on their consistent co-occurring reproduction from sensory stimulation. Both the N1 and P2 peaks can be generated from multiple brain regions, however, they are most consistently produced from within the auditory cortex (Lütkenhöner & Steinsträter, 1998; Näätänen & Picton, 1987; Picton et al., 1974). Additionally, while the N1 and P2 can be generated from all forms of stimuli, they are particularly sensitive to the time/amplitude varying properties of an auditory stimulus. For instance, the amplitude of the N1 and P2 peaks are modulated by variance in sound intensity, with general increases in peak amplitude with louder auditory stimuli (Beagley & Knight, 1967; Keidel & Spreng, 1965).

Another characteristic of auditory stimuli that affects the N1 and P2 is the inter-stimulus intervals between each presentation. With longer periods between stimuli presentation, the N1 and P2 amplitudes increase in magnitude (Budd, Barry, Gordon, Rennie, & Michie, 1998; Davis, Mast, Yoshie, & Zerlin, 1966; Davis & Zerlin, 1966; Keidel & Spreng, 1965). Pitch variation can also modulate the N1/P2 waveforms, with lower frequencies (250-400 Hz) generating higher amplitudes and longer latencies than higher frequency stimuli (1000-3000 Hz) (Antinoro, Skinner, & Jones, 1969; Jacobson, Lombardi, Gibbens, & Ahmad, 1992; Wunderlich & Cone-Wesson, 2001). The above examples concern responses to the exogenous characteristics of stimuli, however, the N1/P2 complex is modulated by endogenous factors. For instance, attentional factors play a role in the size of the N1/P2. Studies show that the level of attention to an auditory stimuli compared to an unattended stimuli will increase the amplitude of the N1 while the amplitude of the P2 is decreased

(Näätänen & Michie, 1979; Näätänen & Picton, 1987). The N1 and P2's generative origin, and their sensitivity to variance in the characteristics of auditory stimuli, outline their usefulness as a functional marker to changes in auditory processing. Because of this, researchers commonly use the N1 and P2 peaks to measure how auditory speech is processed when accompanied by visual speech.

AV Speech Integration at the N1/P2.

There are many studies on AV speech integration that focus specifically on the auditory N1/P2 ERP peaks to identify the neural effect visual speech has on auditory speech processing (Besle et al., 2009; Hisanaga, Sekiyama, Igasaki, & Murayama, 2009; Klucharev et al., 2003; Paris, Kim, & Davis, 2017; Pilling, 2009; Stekelenburg & Vroomen, 2007; Van Wassenhove, Grant, Poeppel, & Halle, 2005; Winneke & Phillips, 2011). This research has outlined two common responses to auditory speech that is presented in conjunction with visual speech cues (AV speech). The first is that visual speech decreases the required demands of processing an auditory speech signal. Visual speech reduces the amplitude of both the N1 and P2 components of an auditory ERP compared to speech from only the acoustic modality (Besle, Fort, Delpuech, & Giard, 2004; Klucharev et al., 2003; Stekelenburg & Vroomen, 2007; Vroomen & Stekelenburg, 2011). The Predictive Coding framework would suggest that visual cues aide auditory processing by generating neuronal expectancy of a stimulus and reducing the prediction error generated by the neuronal populations. The second main finding is that visual speech facilitates the speed of auditory processing. Both the N1 and P2 components occur faster in the waveform when cued by visual speech (Paris et al., 2016b; Van Wassenhove et al., 2005). However, there has been inconsistency between studies on the production of these effects. A number of studies have been unable to reproduce the observed amplitude reduction (Baart & Samuel, 2015) or amplitude latency facilitation (Kaganovich & Schumaker, 2014) for the N1 component and no observed latency facilitation at the P2 (Stekelenburg & Vroomen, 2007).

A large part of these conflicting results may be due to variation in the experimental design of these studies, as mentioned earlier, the N1/P2 are sensitive to the characteristics of the auditory stimuli, differences in, sound intensity (Beagley & Knight, 1967), inter stimulus interval times (Budd et al., 1998) and pitch (Wunderlich & Cone-Wesson, 2001). To address the inconsistent results found within the ERP literature, Baart (2015) quantified the grand average ERP waveforms of multiple studies that measured the N1 and P2 effects of AV integration. He found that the grand average waveforms displayed both the amplitude reduction and speeded latency response of the N1 and P2 for AV ERPs compared to AO ERPs. Despite the occasional inconsistency in the reproduction of these neural modulations, the overall literature indicates that visual speech cues affect both the speed and magnitude of auditory speech processing. It is unclear as to which characteristics of the visual speech cues are providing the predictive information that produces these faciliatory and suppressive effects. Debate has emerged concerning whether the N1/P2 are markers of phonetic AV integration (form cues) or whether non-phonetic information, such as temporal and spatial cues, are producing these responses.

Does the N1/P2 reflect Phonetic or Non-Phonetic AV Integration?

In one of the initial investigations into AV speech integration, Klucharev, Möttönen, and Sams, (2003) attempted to find neural evidence for both phonetic (form) and non-phonetic (temporal and spatial) AV integration. The auditory evoked N1/P2 amplitudes of AV stimuli were compared to auditory-only (AO) and visual-only (VO) stimuli. The main aim was to distinguish between phonetic and non-phonetic integration by comparing ERPs to AV stimuli that were either congruent (auditory /a/ and visual /a/) or incongruent (auditory /a/ and visual

/y/). Importantly, incongruent AV stimuli were chosen to not elicit McGurk type effects as this would be a form of phonetic integration despite AV incongruence. It was found that AV stimuli (for both congruent and incongruent AV) elicited a smaller amplitude at both the N1 and P2 than the combined potentials of AO + VO ERPs (Klucharev et al., 2003). However, no difference in amplitude of the N1/P2 complex between congruent or incongruent AV stimuli was found.

Following these results, Stekelenburg and Vroomen (2007), examined AV integration at the N1/P2 in both multimodal speech and multimodal non-speech perception within multiple contexts. In one experiment, they compared the ERPs to AV speech syllables (/fu/ and /bi/) and AV non-speech stimuli (AV videos of hand clapping and object tapping) against AO and VO stimuli. In their second experiment they compared the ERPs of AV congruent and AV incongruent speech (e.g. auditory /bi/ and visual /fu/) and non-speech stimuli (e.g. auditory clap and visual tap).

Stekelenburg and Vroomen found the N1/P2 was both facilitated and suppressed by non-speech and speech stimuli. Additionally, this facilitation and suppression was the same regardless of the congruency of the AV stimuli. In their third experiment, they compared non-speech AV stimuli with no anticipatory motion (e.g. the tearing of paper) against AO and VO conditions. The main aim was to observe any neural effects at the N1 and P2 when participants could predict the type of sound but not predict when the sound would occur. Despite this manipulation, form predictions did not elicit any effects at the N1 or P2. Both, Klucharev et al. (2003) and Stekelenburg and Vroomen (2007), concluded that the N1/P2 are early markers of non-phonetic AV integration. Suggesting that the temporal and spatial coincidences of AV cues suppressed and facilitated auditory processing but form predictions about what the stimuli will be did not modulate these early neural stages.

Besle, Fort, Delpuech and Giard (2009), argued that based on the experiments of, Klucharev et al. (2003) and Stekelenburg and Vroomen (2007), it cannot be deduced that only spatial and temporal processing of the auditory speech stimuli is occurring at these early N1/P2 stages. Besle et al., (2009) suggest that coming to this conclusion would involve the logical fallacy of converse error. While, the N1/P2 ERPs do produce a waveform morphology that was equivalent for both phonetically congruent and incongruent stimuli, this result does not rule out any sensitivity at the N1/P2 to phonological information.

Indeed, Besle et al, (2009) argue that the visual speech enhancement of early auditory processing is specifically influenced by the phonetic representations in visual speech. Visual speech encodes speech-specific phonetic information onto our Auditory Sensory Memory (ASM). The ASM has been indexed by traditional oddball ERP studies using the resulting Mismatched Negativity (MMN) paradigm to a deviant auditory stimulus. As an example, when participants are presented with a repetitive AV speech stimuli and then presented a deviant visual cue (i.e., a deviant McGurk stimulus), a large MMN deflection generated from the auditory cortex is observed (Colin, 2002). In another study using visual, auditory, and AV deviants a larger MMN to AV deviants was observed in comparison to the sum of both deviant unimodal stimulus (Besle, Fort, & Giard, 2005). These MMN studies suggest that at these early stages AV integration involves the predictive encoding of phonetic cues from visual speech stimuli into ASM. Although to date evidence for this has not been produced conclusively for the N1 or P2.

Another possibility is that the N1 and P2 may have functionally distinct mechanisms. For instance, in a study by Baart, Stekelenburg and Vroomen (2014) it was found that the P2 component may be a marker of speech-specific AV integration. In their design, Sine-Wave Speech (SWS) was used to control if participants perceived auditory speech stimuli as either speech or non-speech (Baart, Stekelenburg, & Vroomen, 2014). SWS is an artificially altered speech stimuli that has been reduced to just the first 3 formants within the speech signal (Remez, Rubin, Pisoni, & Carrell, 1981). Depending on the perceivers experience of SWS, the signal can be heard in either in non-speech mode (as beeps and whistles) or in speech mode. Baart et al., (2014) compared the ERPs to AV congruent SWS and AV incongruent SWS stimuli. When SWS was viewed in speech mode the P2 component was modulated by the congruency of AV cues. This was not found when perceived in non-speech mode. These results suggest that the P2, at least partially, serves as a marker of phonetic binding in AV stimuli.

Paris, Kim and Davis (2017) have also found evidence of the auditory N1's sensitivity to form cues within general AV integration. They presented non-speech AV stimuli in which the preceding visual cues predicted the type of tone that would be heard but gave no indicators to when the sounds would occur. They found that visual form predictions sped up the latency of the N1 response. Missing from this literature is a specific investigation into whether visual information providing phonetic cues without timing or spatial cues, can modulate processing from within the auditory cortex. In Baart, Stekelenburg and Vroomen's (2014) study, the temporal and spatial components of AV speech were not segmented from the SWS cues. Paris, Kim and Davis, (2017) found the influence of form-only AV integration using non-speech stimuli, however, the experiment did not address whether form-only speech-specific phonetic cues can modulate auditory speech perception.

One of the main reasons for this gap in the literature may be to do with how hard it is to separate out visual speech form and timing cues. This difficulty is due to the interconnected nature of the temporal and form properties of visual speech cues (Summerfield, 1992). This is because the articulation of speech syllables over time reveals different information about the phonetic syllable being produced. Thus, it is difficult to completely untangle this crucial timing and form relationship. As mentioned above, we have adopted an approach whereby we use visual stimuli that can convey phonological information but which do not convey inherent timing information. That is, orthography, or print, can offer a functional replacement for one aspect of visual speech. This is because there is already a learned relationship between the symbolic representation of language and their phonetic speech sounds. In a pilot experiment, we examined whether predictive print cues that preceded the presentation of a syllable target facilitated the response times in an auditory identification task, and whether the size of any effect would be similar to that produced by visual speech cues (see Appendix A). Both the printed cue stimuli and visual speech cues facilitated response times compared to control conditions. These experiments validate the use of print stimuli as a tool to investigate the integration of AV speech.

To identify whether speech specific form-only cues can induce facilitation (speeded latency) and suppression (amplitude reduction) of the N1/P2 components we conducted two separate EEG experiments. In the first experiment, we compared the ERPs of AV Speech cues (that provide both timing and form cues) to auditory speech paired with a static face (containing no useful visual predictive cue). In experiment 2, we compared the ERPs to auditory speech when paired with print visual cues that provide either predictive (i.e., printed “ba” preceding auditory /ba/) or non-predictive form information (a meaningless printed control symbol).

Based on the prior literature there are three possible outcomes. 1). If, as Besle et al. (2009) suggest, that visual phonetic content cues influence neuronal processing through sensory memory then we can expect printed cue stimuli that contains predictable content information to modulate the N1 and P2 amplitudes of the auditory ERP. 2). If it is necessary that there is a clear temporal as well as form correspondence between visual and auditory stimuli, then the predictive cues without this property (i.e., the predictive print cues) will have no effect on the ERPs. 3). Finally, it is possible that the N1 or P2 serve functionally distinct roles. If one of the ERP components are uniquely sensitive to the integration of phonetic content cues then we can expect to observe amplitude and latency modulation in that component.

The experimental contrast planned for these experiments differs from that of traditional ERP studies on AV integration (Besle et al., 2009; Paris et al., 2016b; Stekelenburg & Vroomen, 2012, 2015). These studies measure AV integration using an additive model that compares [AV – (AO + VO)] ERPs, the reasoning is to account for possible supra-additive/sub-additive multimodal effects (Barth, Goldberg Brett, & Di, 1995). That is, if the combined effect of unimodal ERPs is equivalent to the ERPs to multimodal stimuli than the effect may not be reflecting actual integrative mechanisms. However, a recent review of these studies (Baart, 2016) highlighted that the additive model design is not essential to observe the N1 and P2 ERP effects of AV integration. Taking this into consideration, the following experiments measure AV integration effects by comparing the N1 and P2 components of AV stimuli against AO stimuli directly. Furthermore, the results of Experiment 1 will provide a baseline to gauge the level of modulation of auditory processing when preceded by the print cues used in Experiment 2.

Experiment 1: The influence of Visual Speech Cues on auditory speech processing

In this experiment, the peak amplitude and peak latency of the N1 and P2 components of auditory ERPs were compared when auditory speech syllables were cued by visual speech (AV, containing both timing and form cues) or Static faces (AO, without any timing or form information). To remove temporal information provided by the initial presentation of a visual stimulus, videos were presented on screen at random time intervals before each stimulus was played. Based on other research (Klucharev et al., 2003; Paris et al., 2016b; Stekelenburg & Vroomen, 2007) it is likely that the predictive visual cues (timing and form) provided in the

AV condition will facilitate and suppress auditory processing at the N1/P2 compared to the AO condition. AV speech will generate a faster latency and reduced amplitude of both the N1 and P2 ERP components than in the AO condition.

Method

Participants

Twenty-five participants (7 men, 18 women) were recruited from Western Sydney University. Participants, aged between 18-30 years old (Mean age = 22.84 years, *SD* = 3.26) were given course credit for participation. Sixteen participants were monolingual native Australian English Speakers, the remaining 9 participants were bilingual with Australian English as their first language. All participants reported having normal hearing, and normal or corrected-to-normal vision. The study was conducted with the approval of the local ethics committee of Western Sydney University. Written consent was obtained from each participant.

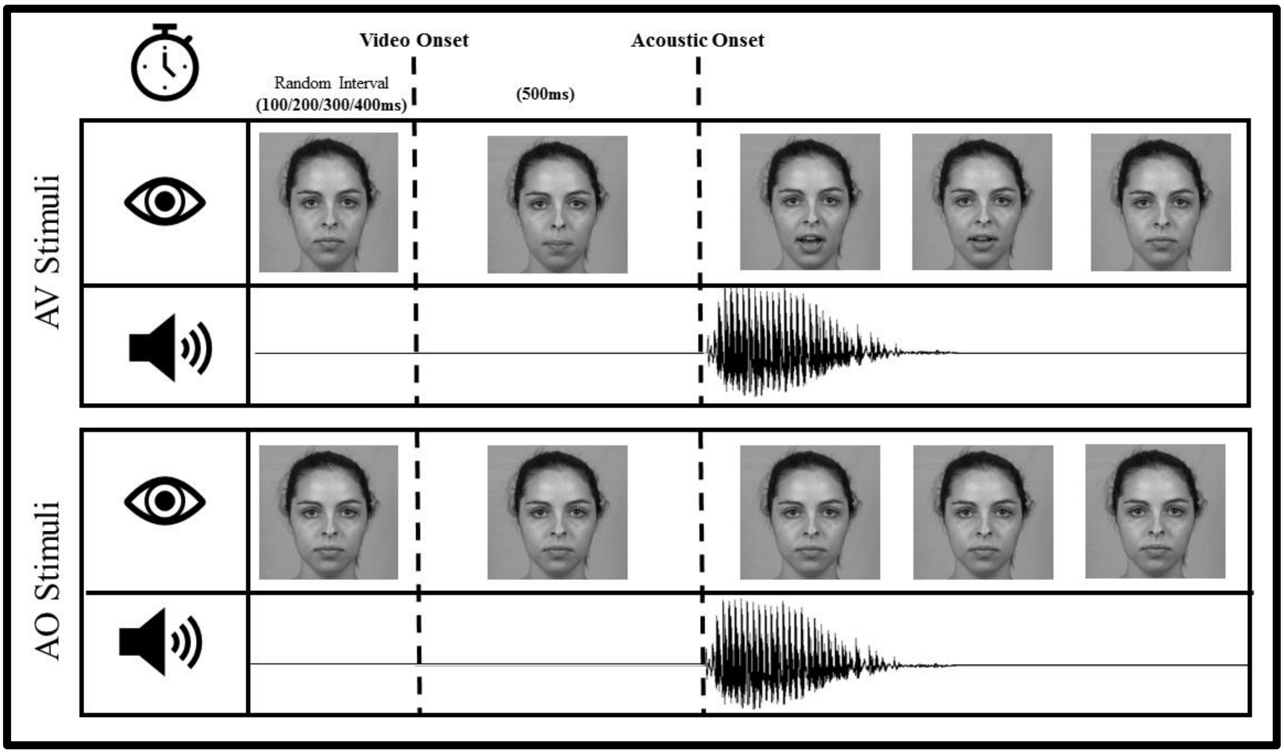
Experimental Design & Stimuli

In this experiment, /ba/ and /da/ speech syllables were presented to participants as AV speech stimuli and AO speech stimuli. To assess that both auditory and visual stimuli were attended to, the experiment also included visual (a red X over speaker’s lips) catch trials and auditory (/ga/ syllables) catch trials (approximately 15% of trials). Participants were tasked to press the spacebar key on the keyboard whenever they heard or saw either catch trial. Speech syllables were produced by both a male and female native speaker of Australian English aged 24-years-old. A total of 480 experimental trials were presented to participants. Sixty /ba/ and 60 /da/ syllables were presented as both AV and AO stimuli.

Different speech stimuli from the pilot experiment (Appendix A) were developed for Experiment 1 and 2. Auditory and visual speech was recorded in a sound attenuated booth. Video was captured with a Sony NXCAM HXR-NX30p camera at 1080p, a 1920x1080 pixel resolution, and 50 fps. Audio was taped with a AT 4033a Transformerless Capacitor Studio microphone at a 44.1khz sample rate. Speakers were video recorded uttering individual /ba/, /da/ and /ga/ syllables; the video captured the speaker's head within frame at a rate of 50fps. Five instances were recorded for each speech syllable, per speaker, equalling a total 30 recorded speech syllables (5 repetitions x 3 syllables x 2 speakers). The duration of auditory syllables ranged from 229 to 473ms (*M* = 315.5, *SD* = 94.2) for /ba/ syllables, 242 to 554ms (*M* = 346.2, *SD* = 118) for /da/ syllables, and 242 – 588ms (*M* = 350, *SD* = 123.1) for /ga/ syllables. Audio intensity was normalised to 70 dB SPL. Video stimuli of each syllable were produced for both AV and AO conditions.

Figure 1.

Timing onsets of the AV and AO stimuli in Experiment 1.



To include all pre-verbal lip and jaw articulation, each stimulus included 500ms of video before and after the onset/offset of auditory voicing. For the AO condition, videos were generated with a static image of the speaker's neutral face taken during recording, which were combined with the auditory speech syllables to create the AO (static) stimuli. All video stimuli were converted into greyscale, at a 500x500 pixel resolution and a frame rate of 25fps. A total of 60 speech stimuli were created (5 repetitions x 3 syllables x 2 speakers x 2 conditions).

Procedure

The experiment took place in an electrically shielded room and participants were individually tested. Visual stimuli were presented on a 16-in CRT monitor that was positioned at eye level, 100 cm directly in front of the participant. Sounds were presented through binaural ER-2A insert earphones (Etymotic research). Videos were displayed using the Psychtoolbox – 3 (Brainard, 1997). Video frames subtended 11° horizontal/vertical visual angle.

A still image of the first frame of the video was presented before each video began. To avoid stimuli beginning at the same time the length of presentation of the initial image frame was randomised over 4 intervals, 100/200/300/400ms (see figure 1). During the experiment, 480 syllable stimuli were presented, /60/ ba and 60 /da/ syllables, in each AV and AO condition, and by each male and female speaker. Stimulus presentation were blocked by speaker, the order of which was counterbalanced, and presentation of syllables and stimulus conditions was randomised. Between two trials a fixation cross was presented for any random duration within 500 to 1500ms. Participants were asked to press the spacebar key whenever they heard or saw an AV or visual catch target (these were presented for 15% of trials, i.e., 80 trials). Catch trials consisted of both 40 AV and 40 visual only targets. AV catch trials consisted of presentation of AV /ga/ targets in both stimuli conditions. For the visual catch target, a red “X” would appear over the talkers’ lips at the onset of the auditory speech syllables. The experiment took approximately 40 minutes to complete.

EEG Recording and analysis preparation

Electrophysiological measurements were recorded with a Biosemi ActiveTwo system (see Biosemi system, Amsterdam, The Netherlands) using an EEG cap with 64 electrodes (10/20 layout – see https://www.biosemi.com/headcap.html). EEG data were recorded at a sampling rate of 256hz. Eight additional electrodes were used: Four ocular electrodes were applied to record eye-blinks and saccades (both vertical and horizontal EOG); two electrodes were applied to the mastoids; and two electrodes were placed as reference during recording (CMS/DRL[1]). A highly-conductive gel was placed between each electrode and scalp to increase conductivity. Prior to cap placement the participants head was brushed to reduce impedances and increase conductivity between the electrodes and scalp (Mahajan & McArthur, 2010)**.**

The EEG events were triggered at the beginning of the audio file, precisely 500ms before the onset of syllable (in analysis this was re-timed to mark the onset of the auditory speech syllables by shifting event triggers by 500ms). Pre-processing and analysis was conducted with EEGLAB version 10 (Delorme & Makeig, 2004). Data were re-referenced

[1] Within the active electrode BioSemi system two electrodes, the Common Mode Sense (CMS) active electrode and the Driven Right Leg (DRL) passive electrode, serve as the "ground" electrodes (traditionally used in bnon-active systems). See https://www.biosemi.com/faq/cms&drl.htm

offline to the mastoid electrodes. EEG data were high pass filtered at 0.15 Hz and low pass filtered at 20 Hz. Data were initially screened for rejection of large artifacts. An automatic channel rejection procedure was conducted such that channels that were 3.5 standard deviations from mean amplitude were removed from the analysis. An Independent Components Analysis (ICA) was conducted and components that revealed stereotypical eye-blinks and saccades were removed. Any rejected channels were interpolated from the remaining electrodes. Data were then epoched at a 1200ms time range, -600ms before and 600ms after auditory onset. Finally, epoched events with average amplitudes surpassing +/-100uV were manually rejected. Peak amplitude and peak latency scores of the ERPs were calculated at electrode Cz, the Midline Central position of the scalp, within time-windows of 70 to 150ms for the N1 and 120 to 250ms for the P2 components. Time windows were based on a visual estimation of the largest and most robust peaks of the ERPs. These time-windows were previously used to measure the early time course of prediction in AV integration (Paris et al., 2016b, 2017; Stekelenburg & Vroomen, 2007).

Results

Data Screening

Data were rejected based on two main criteria. The behavioural accuracy scores for catch trials and the quality of the EEG data. The data from three participants were rejected on the basis of poor accuracy scores (<75% correct). Participants were also screened on the basis of the quality of their EEG data. If >90% of epochs were removed from a data set after the EEG pre-analysis procedures then the data were removed from analysis.

A total of 7 participants were removed due to poor EEG data quality. There were several likely reasons for the high number of participants with poor EEG data. Both the electrode setup and experiment proper involved lengthy procedures (in excess of 1 hour), and this placed demands on the participant's attentional capacity and reduced the conductivity of the electrode gel. For some recordings, the mastoid electrodes were used as reference lost connectivity midway through the experimental procedure. Another factor was participant movement during recording which introduced additional electrical noise. Data were analysed only after all participants were tested, as such, these issues were not immediately apparent to the experimenter. After the screening process, the data from 15 participants were useable for analysis.

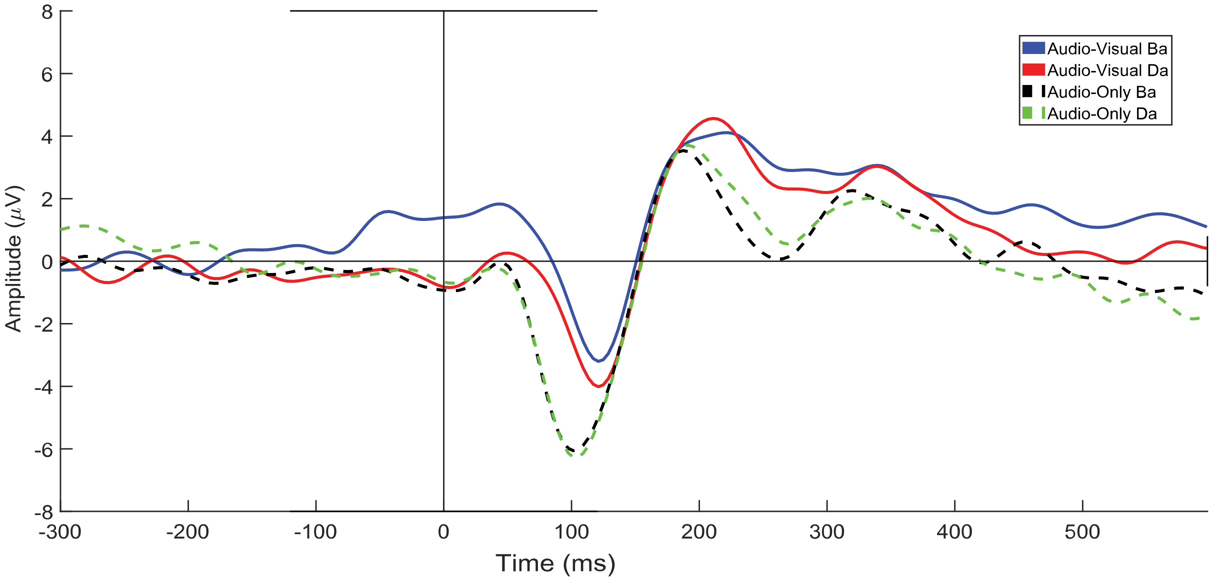
Statistical Analysis

Amplitude and latency data for the N1 and P2 ERP components were analysed for Syllable type (/ba/ or /da/) and for Stimulus condition (Visual Speech, AV versus Static Speech, AO) at electrode Cz. Electrode Cz was selected as it displayed the largest ERP components (Figure 2) and has been used in past studies into AV integration (see Baart, 2016, for a review,). To measure any effect of stimulus presentation on the size of the N1 component, a repeated measures ANOVA was conducted comparing the average peak amplitude of the N1 component for Syllable type x Stimulus condition. There was a main effect of amplitude of stimulus Condition ($F(1,14) = 22.77$, $p = < .01$). Mean amplitude N1 response to AV stimuli was significantly smaller than the mean amplitudes to the AO condition (see Table 1 for a summary of N1 amplitude means). There was no significant effect of Syllable type on the N1 amplitude ($F(1,14) = 2.15$, $p = .16$). Finally, there was no interaction effect between Syllable type and Stimulus condition ($F(1,14) = 2.98$, $p = .38$). To measure any N1 latency effect, an ANOVA comparing Syllable type x Condition for N1 latency scores was conducted. A main effect for Stimulus condition was found ($F(1,14) = 47.2$, $p = <.01$). The average N1 latency was significantly reduced in the AO condition compared to AV one (Table 1). There was no main effect for syllable type ($F(1,14) = .216$, p

= .65) and no significant interaction between syllable type and stimulus condition (*F*(1,14) = .144, *p* = .71).

Figure 2.

Grand Average ERPs as a function of Auditory-Visual Speech and Auditory-Only speech by Syllable type as measured at Cz.



Note: N = 15.

Table 1.

Means and Standard Deviations of N1 peak amplitudes and peak latencies for Experiment 1 as measured at Cz

	Auditory-Visual Speech		Auditory-Only Speech	
	M (*SD*)		M (*SD*)	
Syllable	Amplitude (*µv*)	Latency (*ms*)	Amplitude (*µv*)	Latency (*ms*)
Ba	-1.39(2.81)	119.79(7.29)	-4.49(2.76)	106.38(9.03)
Da	-2.32(2.07)	119.79(8.31)	-4.53(3.08)	104.81(9.01)

To measure any effect of stimulus presentation on the size of the P2 and facilitation, Syllable type x Stimulus condition repeated measure ANOVAs were conducted for P2 amplitude and latency. There was no significant effect of P2 amplitude variation for Stimulus condition (*F*(1,14) = 2.45, *p* = .14) or Syllable type (*F*(1,14) = .00, *p* = .99). Neither was there any significant differences in P2 latency for Stimulus condition (*F*(1,14) = 1.162, *p* = .29) or Syllable type (*F*(1,14) = 0.513, *p* = .486).

Table 2

Means and Standard Deviations of P2 peak amplitudes and peak latencies for Experiment 1 as measures at Cz

	Visual Speech M (*SD*)		Static Speech M (*SD*)	
Syllable	Amplitude (*µv*)	Latency (*ms*)	Amplitude (*µv*)	Latency (*ms*)
Ba	1.56(2.76)	123.57(6.28)	0.37(2.97)	120.83(15.32)
Da	1.32(2.28)	122.26(6.95)	0.61(3.21)	118.75(8.98)

Discussion

It was expected that and increased latency and decreased amplitude of the N1 and P2 components would occur within the AV speech condition relative to the AO condition. We found that the AV speech cues induced a significant reduction in the amplitude of the N1 component. This confirmed our hypothesis that visual speech cueing would reduce auditory processing. However, we did not find the expected increase in latency to AV stimuli over AO stimuli. Instead, the AO condition induced a significantly faster latency score than the AV condition. For the P2 component, neither amplitude or latency scores were significantly different for the AV and AO condition. There was no difference between the amplitude or latencies of /ba/ or /da/ cues at either the N1 or P2. The results are do not fully align with the past research investigating AV integration. These studies have shown that AV speech induces N1 and P2 components with faster latencies and reduced amplitudes than the N1 or P2 components to AO speech (Baart, 2016).

There are several interpretations of these inconsistent findings. For the N1, a suppressive effect was observed with a reduction in amplitude when induced by AV stimuli over AO stimuli. However, mean peak latency was significantly faster in the AO condition than AV. One possible explanation for this seemingly faster result may be due to attention related effects. Past research has indicated that attention has the opposite effect of multisensory prediction on the auditory N1/P2, for instance, the processing of attended stimuli results in a greater amplitude (Näätänen & Michie, 1979; Näätänen & Picton, 1987) and slower latencies (Michalewski, Prasher, & Starr, 1986) of the ERP components than to an unattended stimuli. It is possible that the lack of temporal predictability of the auditory signal in the AO condition, as well as, the extended length between the onset of visual stimuli and the onset of the auditory syllable (ranging 600 – 900*ms*, see figure 1), may have diverted more attention to the auditory signal within the AV condition than in the AO condition. However, in one study that measured the auditory ERPs to either attended, predicted or both attended and predicted stimuli, found that the attentional effect of amplitude enhancement at the N1 occurred only when the sounds were unpredicted (Paris, Kim, & Davis, 2016a). Finally, there was no significant differences of P2 latencies and amplitudes between the AV and AO conditions. This is not unprecedented as both AV facilitation and suppression effects have not always been reproduced: Baart and Samuels (2015) found no amplitude reduction for the N1 and in some studies no latency facilitation was observed for the N1 (Kaganovich & Schumaker, 2014) or for the P2 (Stekelenburg & Vroomen, 2007).

Experiment 2: Predictive Print cues vs. Meaningless Print cues

To identify whether form-only cues can influence the N1 and P2 components of an auditory ERP, we compared print stimuli with predictive content cues versus a print control cue. Like with Experiment 1, the period between the onset of visual cues and the onset of auditory syllables were randomised. If, as Besle, Bertrand, and Giard (2009) suggest, the N1 and P2 components are modulated by the priming of auditory speech via phonetic information than we can expect differences in the peak latencies and peak amplitudes of the

ERP components for predictive print stimuli compared to control print cues. If, the auditory N1 and P2 components are sensitive to only the timing and spatial relationships of AV stimuli then we would observe no differences between the peak amplitudes and peak latencies of the ERP waveforms. Another possibility is that the N1 and P2 components may be functionally distinct, in which case the N1 would display differences in peak amplitude and peak latencies that will not be observed at the P2, or vice versa.

Method

Participants

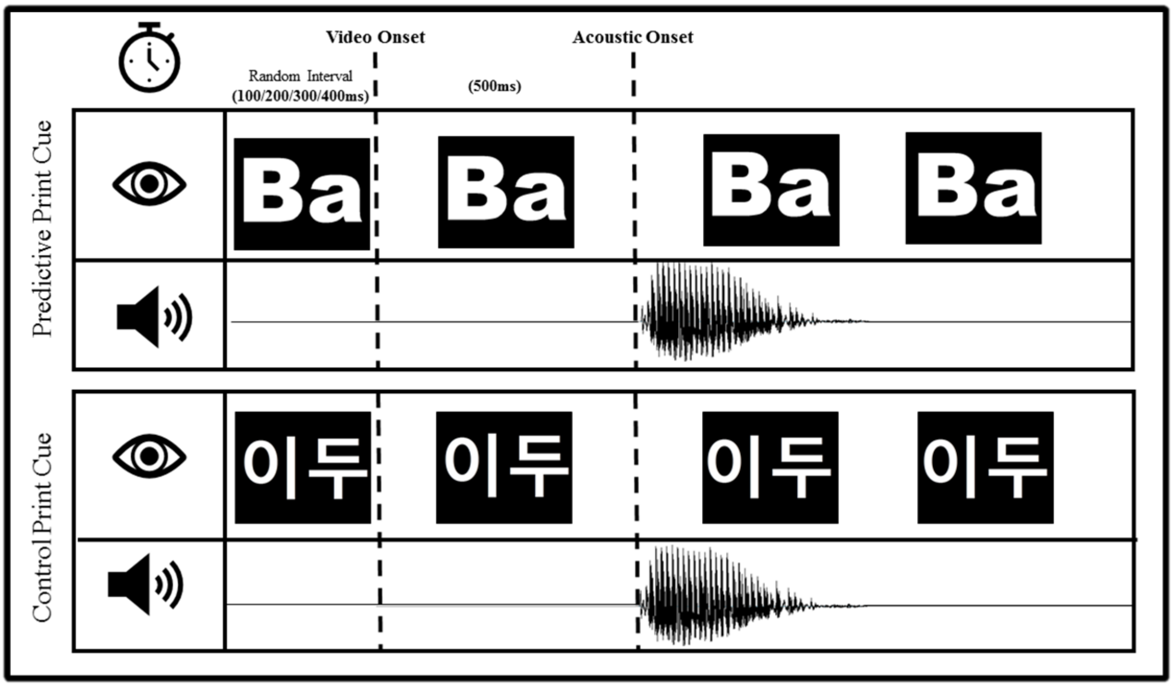
Twenty-Five English-speaking students (10 males, 15 females) were recruited from Western Sydney University. Participant ages ranged from 19 to 47 years, with a mean age of 25.16 years (SD = 5.81). Participants reported having normal hearing, and normal or corrected-to-normal vision. Participants were given course credit for participation and their written consent was obtained. Twelve participants were monolingual native Australian English Speakers, the remaining 13 participants were bilingual with Australian English as their first language. Participants with prior knowledge of or experience with Korean orthography (as this was used as a control condition) were excluded from participation. The study was conducted with the approval of the local ethics committee of Western Sydney University.

Stimuli and Procedure

The same experimental setup as used in Experiment 1 was employed in Experiment 2. The chief difference was that in Experiment 2, /ba/ and /da/ speech syllables were presented to participants paired with either predictive print cues or control print cues. In the Predictive print cue condition, presentations of printed “Ba” and “Da” syllables preceded the auditory syllables. The Control Print condition consisted of presentations of a meaningless print cue, a character from the Korean alphabet (“이두”). To assess that both auditory and visual stimuli were attended to, the experiment also included visual (a red X over print cues in both conditions) catch trials and AV (auditory /ga/ syllables with a visual “Ga” for predictive print cue or visual “이두” for the control) catch trials (approximately 15% of trials). Participants were tasked to press the spacebar key of the keyboard whenever they heard or saw either catch trial. Speech syllables were produced by both a male and female Australian speaker. A total 480 experimental trials were presented to participants. 60 /ba/ and 60 /da/ syllables were presented as both predictive print cues and control print cues.

Figure 3.

Timing onsets of the Predictive Print and Print Control stimuli in Experiment 1.



Printed cue stimuli were developed by constructing 500 x 500 pixel images of meaningful print cues, white English-Roman script (of “Ba”, “Da”, and “Ga” syllables), and the meaningless print cues, characters from Korean alphabet (i.e. “이두”), over a black background. These images were then merged with the audio speech stimuli from Experiment 1 to create AV Print cue video stimuli. Video stimuli were created at 25 fps, with an audio sample rate of 44.1 kHz (70 dB SPL). A total of 30 stimuli were created for each condition.

Within each block there were 120 stimuli for each condition (predictive and control) and within each condition there were 60 presentations of each syllable (/ba/ and /da/). Condition and printed stimuli were randomised within each block. The first image frame for each printed stimulus was presented at a random time interval (100/200/300/400ms) before the video stimuli were played. Participants were asked to press a spacebar whenever infrequent catch trials were presented. A total of 40 catch trials were presented within each block. Catch trials consisted of both auditory and visual stimuli. Auditory catch trials were presentations of the auditory syllable /ga/. For the visual catch trials, a red “X” appeared over the text at the time of auditory onset. The experiment would take approximately 40 minutes to complete.

EEG Recording and analysis preparation

The same procedures in Experiment 1 were utilised in Experiment 2 (see Experiment 1, EEG recording and analysis preparation).

Results

Data Screening

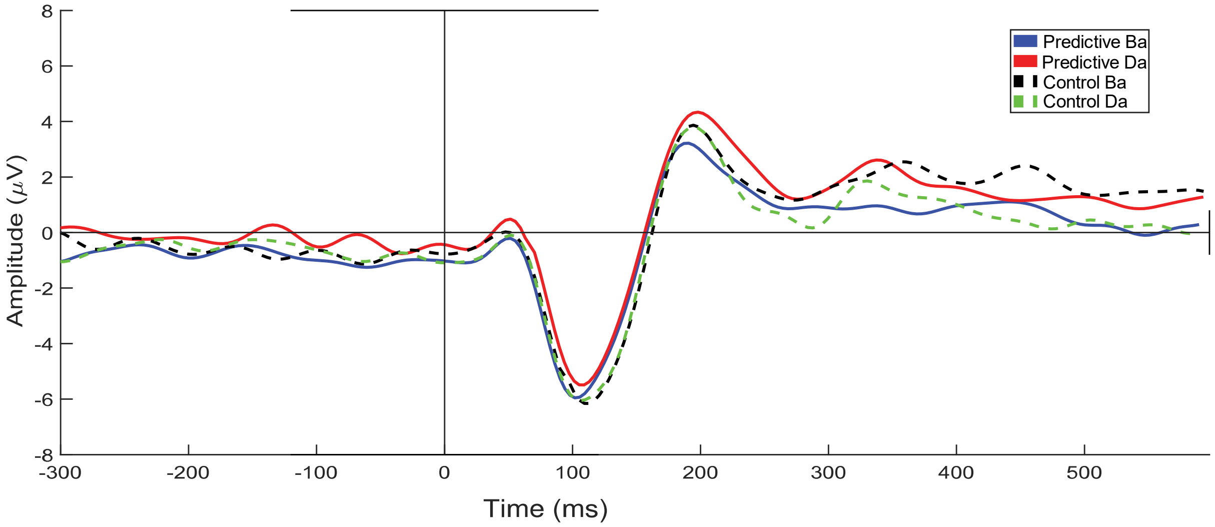
A total of 10 participants data were screened out and hence not analysed. The data from one participant was removed due to a poor overall score on the behavioural task (accuracy less than 75%). The data from the other 9 participants were removed due to the quality of the EEG data recordings. EEG recordings for both experiments were conducted at the same time and similar reasons for the number of participants who produced poor quality data (see Experiment 1 Method) likely applied.

Statistical analyses

As was the case with Experiment 1, changes in ERP amplitude and timing between AO and AV conditions were assessed by examining the amplitude and latency of the peak N1/P2 components. As in Experiemnt 1, analysis was conducted on the averaged data of electrode Cz (Figure 4).

Figure 4.

Grand Average ERPs for Predictive and Control Print cues at electrode Cz.



Note: N = 15.

A repeated measures ANOVA was conducted comparing stimulus condition (predictive or control print cues) x syllable type (/ba/ or /da/) on the N1 amplitude and latency scores. For suppression of negative amplitude there was no effect for stimulus condition (*F*(1,14) = .727, *p* = .408), and no effect for Syllable type (*F*(1,14) = .874, *p* = .366). There were also no main effects found for N1 latency facilitation for stimulus condition (F(1,14) = .742, *p* = .403) or syllable type (*F*(1,14) = .491, *p* = .495). Means for either amplitude or latency scores (as shown in Table 3) did not significantly differ between predictive and control print cues.

Table 3.

Means and Standard Deviations of N1 peak amplitudes and peak latencies for Experiment 2 as measured at Cz

	Predictive print cues		Control print cues	
	Mean (*SD*)		Mean (*SD*)	
Syllable	Amplitude (*μv*)	Latency (*ms*)	Amplitude (*μv*)	Latency (*ms*)
Ba	-4.29(2.51)	109.76(12.64)	-4.38(2.08)	112.86(12.86)
Da	-3.68(2.37)	111.62(9.89)	-4.46(2.50)	113.71(14.41)

Repeated measures ANOVAs were also conducted on the amplitude and latency scores of the P2 component comparing stimulus condition x syllable type. There was no significant effect for P2 amplitude suppression for either stimulus condition ($F(1,14) = 1.39$, $p = .258$) or syllable type ($F(1,14) = 2.09$, .243). Neither was there any significant effect for P2 latency facilitation at stimulus condition ($F(1,14) = 3.14$, p .09) or syllable type ($F(1,14) = 0.5$, $p = .49$).

Table 4.

Means and Standard Deviations of P2 peak amplitudes and peak latencies for Experiment 2 as measured at Cz

	Predictive print cues		Control print cues	
	Mean (*SD*)		Mean (*SD*)	
Syllable	Amplitude (*μv*)	Latency (*ms*)	Amplitude (*μv*)	Latency (*ms*)
Ba	.035(2.59)	119.4(8.69)	.021(2.67)	123.40(19.97)
Da	.976(2.58)	118.46(8.56)	-.1734(2.71)	129.66(30.92)

Discussion

In this experiment, we compared the peak amplitudes and peak latencies, of the N1 and P2 components, to auditory speech preceded by predictive print cues or by control print cues. Importantly, for both stimulus conditions, the duration between visual and auditory stimuli was randomised. Thus, if the N1 and P2 components can be modulated by only phonetic information than we expected significant difference in the peak amplitude/latency scores of predictive print cues versus control cues. On the other hand, if the temporal and spatial characteristics are important in the early neural processing of AV integration, then we are likely to observe no differences at the N1 or P2. Analysis revealed there was no significant difference between either the predictive or control conditions for each component. Neither did the results show any significant differences in amplitude and latency scores between the /ba/ or /da/ syllables. There are multiple interpretations of this finding. The first is that it indicates that cuing with only phonetic information may not be sufficient to modulate auditory processing from within the auditory cortex. An alternative explanation of these findings is that despite the behavioural results of the pilot study (Appendix A), integration of the printed letters and speech sounds does not occur at these early stages of auditory processing.

General Discussion

In this project, we investigated whether only speech-specific phonetic cues provided by print stimuli could induce AV integration at the N1 and P2 of an auditory ERP, then when compared to a control condition. Additionally, we attempted to observe the suppression and facilitation effect on the auditory N1 and P2, that occurs when visual speech cues precede auditory stimuli than by auditory speech alone. We hypothesised three possible outcomes for this project: If AV speech integration is sensitive to speech-specific phonetic form cues then we would observe modulation in the peak amplitudes and peak latencies for both the N1 and P2 ERP components for AV speech stimuli and Predictive print cue stimuli against their AO control cues. On the other hand, if the N1 and P2 components are not sensitive to speech-specific phonetic cues then we expected differences in auditory processing with AV speech against AO speech, but expect no differences in auditory processing between speech preceded by Predictive Print cues against Print Control cues. The third possibility is that the N1 and P2 components differ in their sensitivity to the characteristics of auditory stimuli. If this is the case we may observe modulation of peak amplitude and latencies at either the N1 or P2 components for Predictive Print compared to the Print Control. For the results of Experiment 1, AV speech cues induced a suppression response at the N1 but not at the P2. Additionally, the expected increase in peak latency was not found for AV speech. Instead presentation of visual speech cues generated an N1 component with a slower peak latency than AO speech. In Experiment 2, no differences were found for peak amplitude and peak latencies at both the N1 and P2 for either print cue conditions.

The results of these two EEG experiments lend itself to multiple possible interpretations. In Experiment 1, modulation of auditory processing by AV speech cueing was observed at the N1 level, however, there was no modulatory effect found at the P2. We also confirmed that AV speech reduces the amplitude of the N1 relative to AO speech, however, the latency of the N1 peak was slower for AV stimuli. As mentioned prior, some researchers did not reproduce any modulation effects at the N1 or P2. For instance, Stekelenburg and Vroomen (2007) found no modulatory effects of latency facilitation at the P2. These inconsistencies are usually determined by variance in experimental design. We interpret the slower latency of the N1 and null effects at the P2 to be a possible result of attention related factors. It has been previously observed that attended stimuli slows the latency of the ERP components (Michalewski et al., 1986). In our design AV speech stimuli were the only cues with information relating the onset of the auditory syllable. This may have diverted more attention to the auditory stimulus than in the AO conditions. This attentional difference may account for the latency differences and null effect at the P2.

The results of Experiment 2, indicated that the Predictive Print cues had no influence on auditory speech processing at the N1 or the P2. This null effect allows us to make a few possible conclusions about the nature of AV integration. The neural effects of AV integration at the N1 and P2 are not affected by phonetic cues when there is no temporal or spatial coincidence to the auditory signal. This contradicts the hypothesis proposed by Besle et al. (2009) in which they suggested that the phonetic cues provided by the visual speech signal, attune our ASM to the characteristics of the auditory speech signal. In Experiment 2, the Predictive Print cues provided a clear indicator of the upcoming phonetic auditory signal, however, the results suggest that the information was not encoded within ASM. These results instead, confirm the proposition by, Klucharev et al., (2003) and Stekelenburg and Vroomen, (2007), in which the amplitude reduction and speeded latency of the N1 and P2 are reflective of non-phonetic cueing. It may be possible, that phonetic binding occurs within AV speech perception, but its effectiveness may be contingent on the established temporal and spatial relationship of AV speech.

There are some caveats about the design and implementation of these experiments which may have affected the overall results. The largest complication with this study was the difficulty in obtaining high quality EEG data. Many participants were screened out of data analysis and this high rejection rate reduced the statistical power of the overall experiment. In terms of limitations of the experimental design, the timing difference between onset of visual print cues and the onset of auditory stimuli (random intervals between 600 – 900 ms) could have affected how the visual and auditory stimuli were perceived together. The timing length was purposefully chosen in order to hinder temporal predictability, however, it may have been too long to establish the AV relationship required to induce phonetic binding. Replications of this study should consider varying video presentation within time intervals that are closer to the onset of the natural visual articulation of speech cues, between 100 –

300 ms (Chandrasekaran et al., 2009). This may also reduce the potential attentional effect generated in Experiment 1.

In conclusion we determined that visual phonetic cues contained no temporal or spatial coincidence with the auditory signal did not modulate auditory processing at the N1/P2 ERP peaks. Based on these results it appears that the feedforward multisensory predictions of visual cues may be focused on redirecting the auditory system to when and where stimuli will occur rather than attuning the ASM to what the stimuli will be. This does not entirely discount the possibility of visual phonetic cueing effect on an auditory stimulus, however, its occurrence may be contingent on the temporal and spatial relationship of the AV stimuli. Future avenues of research should focus on exploring the relationship dynamics between phonetic and non-phonetic visual speech cues.

Appendix A

Pilot Experiment

Visual Speech facilitates auditory speech processing, when making auditory identifications participants are significantly faster at responding when cued by AV speech versus AO speech (Paris, Kim, & Davis, 2013). In this pilot experiment, we aimed to investigate whether this behavioural facilitation effect will be equivalently induced by the presentation of visual printed cues. Utilising the paradigm established by Paris, Kim and Davis (2013), we compared the response times to an auditory identification task when the target auditory speech syllables were primed with either visual speech cues (AV condition) versus print cues (Predictive Print Cue condition). The responses to control cues for each experimental condition were also measured. The control cues for the visual speech condition included AO speech paired with a static neutral face. For the print control condition, auditory cues were primed with print stimuli without meaningful information. It is hypothesized that both the AV condition and Predictive Print Cue condition induce faster identification response times than their respective control cues. If the naturally occurring AV speech relationship can facilitate auditory speech identification tasks faster than print cues, we expected a significantly faster identification response time to AV speech. If Predictive Print cues can equivalently facilitate auditory speech processing to that of AV Speech we can expect no significant differences between response times.

Method

Participants.

Thirty-six participants (6 men, 30 women) were recruited from Western Sydney University. Participants, aged between 17 - 37 years old (Mean age = 22.39 years, *SD* = 6.08) were given course credit for participation. All participants reported having normal hearing, and normal or corrected-to-normal vision. All Participants were first-language Australian

English Speakers and had limited to no knowledge of or prior experience with Korean orthography (as this was used as a control condition). The study was conducted with the approval of the local ethics committee of Western Sydney University. Written consent was obtained from each participant.

Experimental Design and Stimuli.

In this experiment, /ba/ and /da/ speech syllables were presented to participants as AV speech stimuli, AO speech stimuli, Predictive Print Cues and Control Print Cues. Participants were asked to identify whether they heard a /ba/ or /da/ syllable. To assess that both auditory and visual stimuli were attended to, the experiment also included visual (a red X over speaker's lips) catch trials and auditory (/ga/ syllables) catch trials (approximately 15% of trials). Participants were tasked to make no response whenever they heard or saw either catch trial. Speech syllables were produced by both a male and female Australian speaker. A total 960 experimental trials were presented to participants. These trials were divided by which speaker produced each syllable (480 trials per speaker) and further divided by each condition (120 for each condition within speaker). Within these divisions 60 /ba/ and 60 /da/ syllables were presented.

Auditory and visual speech was recorded from the male and female 23-year-old Australian speakers in a sound attenuated booth. Video was recorded with a Sony NXCAM HXR-NX30p video camera (1920 x 1080 full HD, at a rate of 50 fps) and audio was recorded with a AT 4033a Transformerless Capacitor Studio microphone (at a sampling rate of 44.1khz) were video recorded uttering individual /ba/, /da/ and /ga/ syllables; the video captured the speaker's head within frame at a rate of 50fps. 10 instances were recorded for each speech syllable (3), per speaker (2), equalling a total 60 recorded speech syllables. Audio intensity was normalised to 60 dB SPL.

Video stimuli of each syllable were produced for each experimental condition. To include all pre-verbal lip and jaw articulation, each stimulus included 500 ms of video before and after the onset/offset of auditory voicing. The AV condition contained both the video and audio of each speaker's syllable. For the AO condition, videos were generated with a static image of the speaker's neutral face taken during recording, which were combined with the auditory speech syllables to create the AO (static) stimuli. To create Predictive Print cue stimuli images of white English-roman script of "Ba", "Da", and "Ga" syllables on a black background were converted into videos and synchronised with their respective auditory speech syllable. For the Control Print cues a symbol from the Korean language system ("이두") was converted into a video (white lettering over a black background) and synchronised with each speech syllable. All stimuli were created with a 500x500 pixel resolution at a frame rate of 25fps. 60 stimuli were created (10 repetitions x 3 syllables x 2 speakers) for each condition, totalling 240 speech stimuli overall. Additionally, Visual catch trials were created for each /ba/ and /da/ stimuli. These consisted of a red X appearing over the speakers lips or over the printed lettering at 500 ms into each video (time of auditory onset).

Procedure.

Participants were tested individually in a sound-attenuated booth. Video was presented on a 17" LCD monitor and audio through overear Koss UR-20 headphones. Trials were blocked by speaker (male and female) and Cue type (Printed or Visual cues), presentation of each block was counterbalanced. Stimuli were presented with the psychophysics display software, DMDX (Forster & Forster, 2003). Within each block syllable trials were randomised. Each trial was presented one at a time. A fixation cross was presented for 1 second between each trial. Participants were asked to identify ("as quickly and as accurately as possible") whether they heard a "ba" or "da" syllable with either a left or right button push on a 2-button response box. Left/Right button responses were counterbalanced between each participant. Participants were also asked to not make a response whenever they heard or saw an infrequent visual or AV catch trials (15% of trials).

Results

Data Screening.

Participants were screened based on two measures, poor accuracy of identification responses (<75% correct responses) and outlying reaction times (3 standard deviations from the mean). Two participants were removed due to poor accuracy scores and another two participants were removed based on outlying reaction times. A remaining 32 participants were available for analysis.

Statistical Analysis.

A 2x2 repeated measures ANOVA was conducted comparing the mean reaction times of cue type (Face cue versus Print cue) x the modality type (AV experimental condition vs AO control conditions). There was a main effect of Modality Type ($F(1,31) = 233.03$, $p = < .01$) with both AV speech and Predictive print stimuli having significantly faster mean response times than their respective controls (see Table 5). There was no significant difference between Cue Type ($F(1,32) = 3.57$, $p = .06$), although there appears to be a trend towards AV speech cues inducing faster identification response times then the other experimental conditions (Table 5). Additionally, based on a visual inspection of AV speech scores there appears to be a larger mean variance than in the other conditions. Finally, There was no interaction effect between Visual Cue type x Modality Cue type ($F(1,31) = 71.46$, $p = .692$).

Table 5.

Means and Standard deviations of auditory identification reaction times (RT) to each condition.

	Visual Speech Cues (RT)	Printed Cues (RT)
	M (*SD*)	M (*SD*)
AV Exp.	9993(141.61)	1015.45(133.07)
AO Control	1066.22(123.17)	1091.18(134.46)

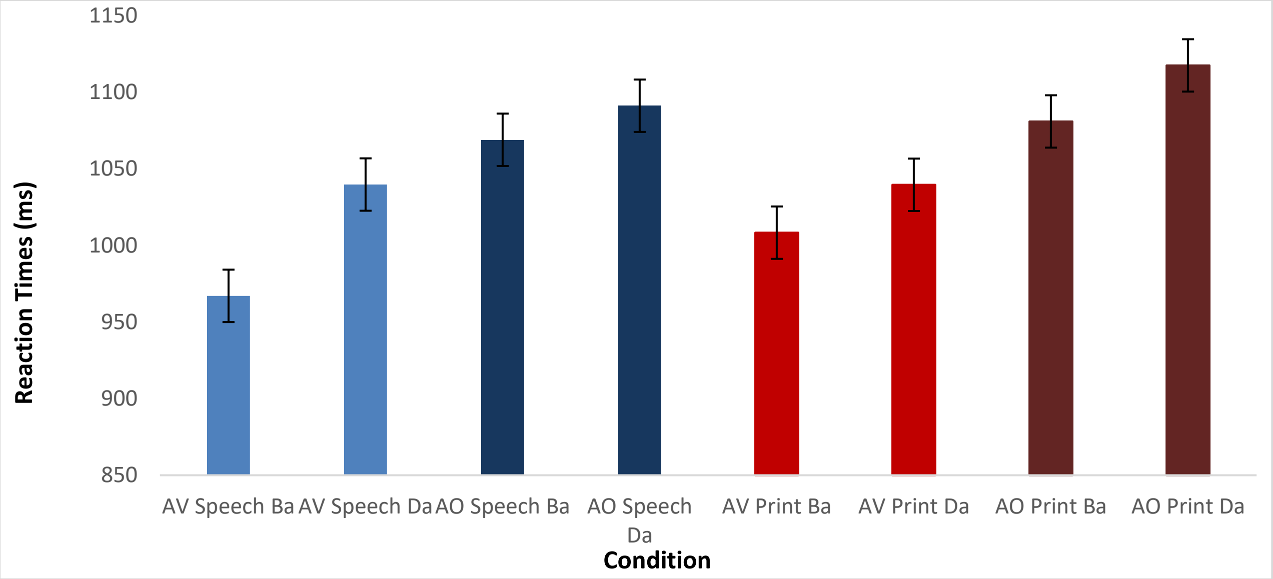
Note: RT in milliseconds

Post-hoc Analysis

Based on the larger observed variance between mean response times within the AV speech condition and additional repeated measures ANOVA was conducted comparing Syllable Type (ba or da syllables) x Cue Type x Modality Type. This revealed a main effect of Syllable Type (*F*(1,31) = 24.48, *p* = <.01), Ba syllables generated faster mean response times than Da syllables.

Figure 5.

Mean reaction times of auditory identification of /ba/ and /da/ syllables in each condition



Note: Error bars represent the standard error

There was still a main effect for modality type, *F*(,1,31) = 240.45, *p* < .01. No significant effects remained with cue type, *F*(1,31) = 2.87, *p* = 0.1. There was also a significant effect between syllable type and condition type, *F*(1,31) = 9623.55 *p* = .01. Finally, a significant effect was also found between syllable type, cue type and condition type, *F*(1,31) = 10.67, *p* = <.01. /ba/ syllables induced faster identification responses than /da/ syllables across all conditions (figure 5).

Discussion

This pilot study investigated whether visual print cues can facilitate auditory speech processing at an equivalent rate to visual speech cues. We compared the reaction times of AV speech cues with AV print cues and with their respective AO control cues on a speech identification task. It is hypothesized that both the AV condition and Predictive Print Cue condition will induce faster identification response times than their respective control cues. If the naturally occurring AV speech relationship can facilitate auditory speech identification tasks faster than print cues, we expected a significantly faster identification response time to AV speech. However, if visual Print cues can equivalently facilitate auditory speech processing to that of AV Speech we can expect no significant differences between response times.

It was initially found that both AV speech cues and AV print cues produced faster mean reaction times than AO control cues. This suggested that visual speech and visual print cues can facilitate auditory speech processing at an equivalent rate. However, when we conducted post-hoc analysis we found significant interaction effects between Syllable Type, Cue Type and Condition Type. /ba/ syllables induced faster response times than /da/ syllables overall, however, this variance was much larger within the AV speech condition. This could be accounted for by the differences in the production of /ba/ and /da/ syllables. Another interpretation is that the audio-visual catch trial increased the difficulty when responding to

/da/ syllables over /ba/ syllables. Overall, in this study we identified that both printed cue stimuli and visual speech stimuli facilitate auditory identification response times.

References

Antinoro, F., Skinner, P. H., & Jones, J. J. (1969). Relation between Sound Intensity and Amplitude of the AER at Different Stimulus Frequencies. *The Journal of the Acoustical Society of America, 46*(6B), 1433–1436. https://doi.org/10.1121/1.1911881

Baart, M. (2016). Quantifying lip-read-induced suppression and facilitation of the auditory N1 and P2 reveals peak enhancements and delays. *Psychophysiology, 53*(9), 1295–1306. https://doi.org/10.1111/psyp.12683

Baart, M., & Samuel, A. G. (2015). Turning a blind eye to the lexicon: ERPs show no cross-talk between lip-read and lexical context during speech sound processing. *Journal of Memory and Language*, *85*, 42–59. https://doi.org/10.1016/j.jml.2015.06.008

Baart, M., Stekelenburg, J. J., & Vroomen, J. (2014). Electrophysiological evidence for speech-specific audiovisual integration. *Neuropsychologia*, *53*(1), 115–121. https://doi.org/10.1016/j.neuropsychologia.2013.11.011

Beagley, H. A., & Knight, J. J. (1967). Changes in Auditory Evoked Response with Intensity. *The Journal of Laryngology & Otology*, *81*(8), 861–873. https://doi.org/10.1017/S0022215100067815

Besle, J., Bertrand, O., & Giard, M. H. (2009). Electrophysiological (EEG, sEEG, MEG) evidence for multiple audiovisual interactions in the human auditory cortex. *Hearing Research*, *258*(1–2), 143–151. https://doi.org/10.1016/j.heares.2009.06.016

Besle, J., Fort, A., Delpuech, C., & Giard, M. H. (2004). Bimodal speech: Early suppressive visual effects in human auditory cortex. *European Journal of Neuroscience*, *20*(8), 2225–2234. https://doi.org/10.1111/j.1460-9568.2004.03670.x

Besle, J., Fort, A., & Giard, M.-H. (2005). Is the auditory sensory memory sensitive to visual information? *Experimental Brain Research. Experimentelle Hirnforschung. Experimentation Cerebrale*, *166*(3–4), 337–344. https://doi.org/10.1007/s00221-005-2375-x

Brainard, D. H. (1997). The Psychophysics Toolbox. *Spatial Vision*, *10*, 433–436. https://doi.org/10.1163/156856897X00357

Budd, T. ., Barry, R. J., Gordon, E., Rennie, C., & Michie, P. . (1998). Decrement of the N1 auditory event-related potential with stimulus repetition: habituation vs. refractoriness. *International Journal of Psychophysiology*, *31*(1), 51–68. https://doi.org/10.1016/S0167-8760(98)00040-3

Callan, D. E., Jones, J. a, Munhall, K., Kroos, C., Callan, A. M., & Vatikiotis-Bateson, E. (2004). Multisensory integration sites identified by perception of spatial wavelet filtered visual speech gesture information. *Journal of Cognitive Neuroscience*, *16*(5), 805–816. https://doi.org/10.1162/089892904970771

Chandrasekaran, C., Trubanova, A., Stillittano, S., Caplier, A., & Ghazanfar, A. A. (2009). The natural statistics of audiovisual speech. *PLoS Computational Biology*, *5*(7). https://doi.org/10.1371/journal.pcbi.1000436

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, *36*(3), 181–204. https://doi.org/10.1017/S0140525X12000477

Colin, C. (2002). Mismatch negativity evoked by the McGurk–MacDonald effect: a phonetic representation within short-term memory. *Clinical Neurophysiology*, *113*(4), 495–506. https://doi.org/10.1016/S1388-2457(02)00024-X

Davis, H., Mast, T., Yoshie, N., & Zerlin, S. (1966). The slow response of the human cortex to auditory stimuli: Recovery process. *Electroencephalography and Clinical Neurophysiology*, *21*(2), 105–113. https://doi.org/10.1016/0013-4694(66)90118-0

Davis, H., & Zerlin, S. (1966). Acoustic Relations of the Human Vertex Potential. *The Journal of the Acoustical Society of America*, *39*(1), 109–116. https://doi.org/10.1121/1.1909858

Forster, K. I., & Forster, J. C. (2003). DMDX: A Windows display program with millisecond accuracy. *Behavior Research Methods, Instruments, & Computers*, *35*(1), 116–124. https://doi.org/10.3758/BF03195503

Friston, K. (2005). A theory of cortical responses. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *360*(1456), 815–36. https://doi.org/10.1098/rstb.2005.1622

Giard, M. H., & Peronnet, F. (1999). Auditory-Visual Integration during Multimodal Object Recognition in Humans: A Behavioral and Electrophysiological Study. *Journal of Cognitive Neuroscience*, *11*(5), 473–490. Retrieved from ezproxy.uws.edu.au/login?url=http://go.galegroup.com/ps/i.do?p=AONE&sw=w&u=uwsydney&v=2.1&it=r&id=GALE%7CA57815809&asid=3343d005b2727fd97e6326c40a2bbd26

Glover, G. H. (2011). Overview of functional magnetic resonance imaging. *Neurosurg Clin N Am*, *22*(2), 133–139. https://doi.org/10.1016/j.nec.2010.11.001.Overview

Hisanaga, S., Sekiyama, K., Igasaki, T., & Murayama, N. (2009). Audiovisual speech perception in Japanese and English: inter-language differences examined by event-related potentials. *Avsp*, 38–42.

Jacobson, Lombardi, Gibbens, Ahmad, N. (1992). The effects of stimulus frequency and recording site on the amplitude and latency of multichannel cortical auditory evoked potential (CAEP) component N1. *Ear and Hearing*, *13*(5), 300–306.

Kaganovich, N., & Schumaker, J. (2014). Audiovisual integration for speech during mid-childhood: Electrophysiological evidence. *Brain and Language*, *139*, 36–48. https://doi.org/10.1016/j.bandl.2014.09.011

Keidel, W. D., & Spreng, M. (1965). Audiometric Aspects and Multisensory Power-Functions of Electronically Averaged Slow Evoked Cortical Responses in Man. *Acta Oto-Laryngologica*, *59*(2–6), 201–210. https://doi.org/10.3109/00016486509124553

Klucharev, V., Möttönen, R., & Sams, M. (2003). Electrophysiological indicators of phonetic and non-phonetic multisensory interactions during audiovisual speech perception. *Cognitive Brain Research*, *18*(1), 65–75. https://doi.org/10.1016/j.cogbrainres.2003.09.004

Lütkenhöner, B., & Steinsträter, O. (1998). High-precision neuromagnetic study of the functional organization of the human auditory cortex. *Audiology & Neurotology*, *3*(2–3), 191–213. Retrieved from http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=9575385

MacDonald, J., & McGurk, H. (1978). Visual influences on speech perception processes. *Perception & Psychophysics*, *24*(3), 253–257. https://doi.org/10.3758/BF03206096

Mahajan, Y., & McArthur, G. (2010). Does combing the scalp reduce scalp electrode impedances? *Journal of Neuroscience Methods*, *188*(2), 287–289. https://doi.org/10.1016/j.jneumeth.2010.02.024

Meredith, A. M., & Stein, B. E. (1986). Visual, auditory, and somatosensory convergence on cells in superior colliculus results in multisensory integration. *Journal of Neurophysiology*, *56*(3), 640–662.

Michalewski, H. ., Prasher, D. ., & Starr, A. (1986). Latency variability and temporal interrelationships of the auditory event-related potentials (N1, P2, N2, and P3) in normal subjects. *Electroencephalography and Clinical Neurophysiology/Evoked Potentials Section*, *65*(1), 59–71. https://doi.org/10.1016/0168-5597(86)90037-7

Murali, S., & Kulish, V. V. (2007). Modeling of evoked potentials of electroencephalograms: An overview. *Digital Signal Processing: A Review Journal*, *17*(3), 665–674. https://doi.org/10.1016/j.dsp.2006.09.004

Näätänen, R., & Michie, P. T. (1979). Early selective-attention effects on the evoked potential: A critical review and reinterpretation. *Biological Psychology*, *8*(2), 81–136. https://doi.org/10.1016/0301-0511(79)90053-X

Näätänen, R., & Picton, T. (1987). The N1 Wave of the Human Electric and Magnetic Response to Sound: A Review and an Analysis of the Component Structure. *Psychophysiology*, *24*(4), 375–425. https://doi.org/10.1111/j.1469-8986.1987.tb00311.x

Paris, T., Kim, J., & Davis, C. (2013). Visual speech form influences the speed of auditory speech processing. *Brain and Language*, *126*(3), 350–356. https://doi.org/10.1016/j.bandl.2013.06.008

Paris, T., Kim, J., & Davis, C. (2016a). The Processing of Attended and Predicted Sounds in Time. *Journal of Cognitive Neuroscience*, *28*(1), 158–165. https://doi.org/10.1162/jocn_a_00885

Paris, T., Kim, J., & Davis, C. (2016b). Using EEG and stimulus context to probe the modelling of auditory-visual speech. *Cortex*, *75*, 220–230. https://doi.org/10.1016/j.cortex.2015.03.010

Paris, T., Kim, J., & Davis, C. (2017). Visual form predictions facilitate auditory processing at the N1. *Neuroscience*, *343*, 157–164. https://doi.org/10.1016/j.neuroscience.2016.09.023

Picton, T. W., Hillyard, S., Krausz, H., & Galambos, R. (1974). Human auditory evoked potentials. I: Evaluation of components. *Electroencephalography and Clinical Neurophysiology*, *36*, 179–190. https://doi.org/10.1016/0013-4694(74)90155-2

Pilling, M. (2009). Auditory event-related potentials (ERPs) in audiovisual speech perception. *Journal of Speech, Language, and Hearing Research : JSLHR*, *52*(4), 1073–81. https://doi.org/10.1044/1092-4388(2009/07-0276)

Rao, R. P., & Ballard, D. H. (1999). Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive field effects. *Nature Neuroscience*, *2*(1), 79–87.

Remez, R., Rubin, P., Pisoni, D., & Carrell, T. (1981). Speech perception without traditional speech cues. *Science*, *212*(4497), 947–949. https://doi.org/10.1126/science.7233191

Sejnowski, T. J., Churchland, P. S., & Movshon, J. A. (2014). Putting big data to good use in neuroscience. *Nature Neuroscience*, *17*(11), 1440–1. https://doi.org/10.1038/nn.3839

Stein, B. E., Burr, D., Constantinidis, C., Laurienti, P. J., Alex Meredith, M., Perrault, T. J., … Lewkowicz, D. J. (2010). Semantic confusion regarding the development of multisensory integration: a practical solution. *European Journal of Neuroscience*, *31*(10), 1713–1720. https://doi.org/10.1111/j.1460-9568.2010.07206.x

Stekelenburg, J. J., & Vroomen, J. (2007). Neural correlates of multisensory integration of ecologically valid audiovisual events. *Journal of Cognitive Neuroscience*, *19*(12), 1964–1973. https://doi.org/10.1162/jocn.2007.91213

Stekelenburg, J. J., & Vroomen, J. (2012). Electrophysiological correlates of predictive coding of auditory location in the perception of natural audiovisual events. *Frontiers in Integrative Neuroscience*, *6*(May), 26. https://doi.org/10.3389/fnint.2012.00026

Stekelenburg, J. J., & Vroomen, J. (2015). Predictive coding of visual-auditory and motor-auditory events: An electrophysiological study. *Brain Research*, *1626*, 88–96. https://doi.org/10.1016/j.brainres.2015.01.036

Sumby, W. H., & Pollack, I. (1954). Visual Contribution to Speech Intelligibility in Noise. *The Journal of the Acoustical Society of America*, *26*(2), 212–215. https://doi.org/10.1121/1.1907309

Summerfield, Q. (1992). Lipreading and audio-visual speech perception. *Philosophical Transactions: Biological Sciences*, *335*(1273), 71–78. https://doi.org/10.1098/rstb.1992.0009

Van Wassenhove, V., Grant, K. W., Poeppel, D., & Halle, M. (2005). Visual Speech Speeds up the Neural Processing of Auditory Speech. *Source: Proceedings of the National Academy of Sciences of the United States of America*, *102*(4), 1181–1186. https://doi.org/10.1073/pnas.0408949102

Venezia, J. H., Thurman, S. M., Matchin, W., George, S. E., & Hickok, G. (2016). Timing in audiovisual speech perception: A mini review and new psychophysical data. *Attention, Perception, & Psychophysics*, *78*(2), 583–601. https://doi.org/10.3758/s13414-015-1026-y

Vroomen, J., & Stekelenburg, J. J. (2011). Perception of intersensory synchrony in audiovisual speech: Not that special. *Cognition*, *118*(1), 78–86. https://doi.org/10.1016/j.cognition.2010.10.002

Winneke, A. H., & Phillips, N. A. (2011). Does audiovisual speech offer a fountain of youth for old ears? An event-related brain potential study of age differences in audiovisual speech perception. *Psychol Aging, 26*(2), 427–438. https://doi.org/10.1037/a0021683

Wolpaw, J. R., & Penry, J.K. (1975). A temporal component of the auditory evoked response. *Electroencephalography and clinical neurophysiology, 39*(6), 609-620. http://dx.doi.org/10.1016/0013-4694(75)90073-5

Wunderlich, J. L., & Cone-Wesson, B. K. (2001). Effects of stimulus frequency and complexity on the mismatch negativity and other components of the cortical auditory-evoked potential. *The Journal of the Acoustical Society of America*, *109*(4), 1526–1537. https://doi.org/10.1121/1.1349184