The Cortical Representation of the Speech Envelope is Earlier for Audiovisual Speech than Audio Speech

Running Head: Earlier Representation of Continuous Audiovisual Speech

Michael J. Crosse,1,2 Edmund C. Lalor 1,2,3†

1School of Engineering, 2Trinity Centre for Bioengineering and 3Trinity College Institute of Neuroscience, Trinity College Dublin, Dublin 2, Ireland.

Author contributions: M.J.C. and E.C.L. designed the experiment. M.J.C. collected and analyzed the data. M.J.C. and E.C.L. wrote the manuscript.

Keywords: multisensory integration, analysis-by-synthesis, latency, EEG, TRF

Number of pages: 28
Number of figures: 6, Number of tables: 0
Number of words: Abstract – 224, Introduction – 814, Discussion – 2093

Conflict of Interest: None

† Correspondence:
Edmund C. Lalor, Ph.D.
Trinity College Dublin
Dublin 2, Ireland
Vox: +353 1 8961743
Fax : +353 1 6772442
e-mail : edlalor@tcd.ie

Copyright © 2014 by the American Physiological Society.
ABSTRACT

Visual speech can greatly enhance a listener’s comprehension of auditory speech when they are presented simultaneously. Efforts to determine the neural underpinnings of this phenomenon have been hampered by the limited temporal resolution of hemodynamic imaging and the fact that electro- and magnetoencephalographic (EEG/MEG) data are usually analyzed in response to simple, discrete stimuli. Recent research has shown that neuronal activity in human auditory cortex tracks the envelope of natural speech. Here, we exploit this finding by estimating a linear forward mapping between the speech envelope and EEG data and show that the latency at which the envelope of natural speech is represented in cortex is shortened by >10 ms when continuous audiovisual speech is presented in comparison to audio-only speech. In addition, we use a reverse mapping approach to reconstruct an estimate of the speech stimulus from the EEG data and, by comparing the bimodal estimate to the sum of the unimodal estimates, find no evidence of any nonlinear additive effects in the audiovisual speech condition. These findings point to an underlying mechanism that could account for enhanced comprehension during audiovisual speech. Specifically, we hypothesize that low-level acoustic features that are temporally coherent with the preceding visual stream may be synthesized into a speech object at an earlier latency which may provide an extended period of low-level processing before extraction of semantic information.
INTRODUCTION

During natural everyday conversation, we routinely process speech using both our auditory and visual systems. Indeed visual information can improve speech comprehension in both noisy (Sumby and Pollack 1954) and noise-free environments (Reisberg et al. 1987). This perceptual improvement may be driven largely by our ability to utilize the temporal precedence of visual speech to predict incoming auditory information (Arnal et al. 2009; Besle et al. 2004; van Wassenhove et al. 2005). Although we easily integrate continuous audio and visual speech, the temporal delay between the unimodal streams varies with speaker and content by up to several hundred milliseconds (Chandrasekaran et al. 2009), making it an intrinsically complex process. Examining the temporal dynamics of audiovisual speech processing is therefore essential to understanding how information from the two modalities is integrated into one perceptual representation.

The limited temporal resolution of functional magnetic resonance imaging (fMRI) makes it unsuitable for investigating the rapid dynamics of ongoing speech. In contrast, EEG and MEG offer excellent temporal resolution, rendering them more suitable methodologies for addressing this subject. However, the standard technique for investigating sensory processing is to repeatedly present discrete stimuli and to derive an average response. In the context of speech, this has largely constrained researchers to examining the processing of individual syllables. This approach overlooks the complex dynamics of speech which is continuous in nature and severely constrains the stimuli in terms of their semantic content. However, within this context, many studies have shown important effects of visual input including the attenuation of key auditory processing indices (Arnal et al. 2009; Besle et al. 2004; Klucharev et al. 2003; Pilling 2009; Stekelenburg and Vroomen 2007; van Wassenhove et al. 2005).
Several of these studies have also reported shorter auditory response latencies when auditory and visual syllables are presented concurrently (Arnal et al. 2009; Stekelenburg and Vroomen 2007; van Wassenhove et al. 2005). This effect has been shown not just for congruent audiovisual syllables, but also for discrete non-speech stimuli (Stekelenburg and Vroomen 2007), discrete incongruent speech stimuli (Arnal et al. 2009; Stekelenburg and Vroomen 2007) and discrete monkey vocalizations (Chandrasekaran et al. 2013). However, to date no evidence has been advanced for such a shift in the case of natural continuous speech in humans.

While the latency effects reported to discrete stimuli fit with the fact that visual speech precedes auditory speech (Chandrasekaran et al. 2009), these effects can complicate the issue of examining nonlinear additive effects between the responses evoked by different modes of speech, such as the attenuation effects mentioned above. This is because the dominant components of event-related potentials (ERPs) are generated by multiple temporally overlapping sources and, hence are likely comprised of several subcomponents (Di Russo et al. 2005). A shift in the latency of any one of these contributions can result in both latency and amplitude changes in the measured ERP component. Indeed, latency shifts in the ERP components can also have an impact on temporally adjacent ERP components, affecting their amplitude (Luck 2004). The reverse is also true in that amplitude differences between components or subcomponents could masquerade as latency shifts when examining ERPs; an issue which has been raised before (Stekelenburg and Vroomen 2007). New approaches are necessary to disentangle latency effects from amplitude effects in human AV speech processing.

This study seeks to investigate this issue using two complementary methods that exploit the fact that changes in cortical activity track the dynamic changes in the speech stimulus (Lalor and Foxe 2010; Millman et al. 2013; Nourski et al. 2009). Specifically we examine the
effects of visual speech on linear response function measures that index how the amplitude envelope of natural speech maps forward to the recorded EEG (Lalor and Foxe 2010). We also exploit a reverse mapping approach between the EEG and the amplitude envelope (Mesgarani et al. 2009) to quantify multisensory amplitude effects in a manner that is relatively free from the confounds of latency effects. In addition to allowing us to examine continuous AV speech processing, our assumption of a linear relationship between the envelope and EEG data has two methodological benefits. Firstly, it minimizes the contribution of the visual stimulus during the timeframe of the auditory response because, unlike discrete stimuli, the dynamics of the visual stimulus are not tightly time-locked to those of the audio stimulus. As such, we can compare neuronal responses to audio-only speech with those from audiovisual speech in terms of how just the auditory stream was represented. Secondly, unlike the ERP, it produces response measures that are indicative of feedforward activity in early auditory cortex (Lalor et al. 2009). We anticipate that these two properties of the approach will allow us to search for latency effects in auditory cortex in the context of continuous stimuli and to provide greater clarity on the issue of latency shifts versus amplitude effects than can be gleaned from examining ERPs derived from discrete stimuli.

METHODS

Subjects

Fourteen right-handed volunteers aged between 20 and 35 years participated with informed consent (4 females; mean age, 25.4 years). All subjects were free of neurological diseases, had self-reported normal hearing and normal or corrected-to-normal vision. The experiment was undertaken in accordance with the Declaration of Helsinki and was approved by the Ethics Committee of the Health Sciences Faculty at Trinity College Dublin, Ireland.
**Stimuli and experimental procedure**

The stimuli were drawn from a collection of videos featuring a native speaker of English (male). The speaker’s face was positioned at the centre of the camera shot which alternated between medium close-up and close-up while the speaker addressed the camera at all times. There was no background movement or noise. Using VideoPad Video Editor (NCH Software), seven videos were truncated to 120 s in length and rendered into 640 × 360 pixel movies with a digitization rate of 29.97 frames/s (1 frame = 33.37 ms). The stereo soundtracks were digitized at 44.1 kHz with 16-bit resolution. Stimulus presentation was conducted using software by Presentation (Neurobehavioral Systems) and delivered using a 19” CRT monitor and Sennheiser HD650 headphones. Each of the seven videos was presented in audio-only (A), visual-only (V) and audiovisual (AV) format, the order of which was randomized between blocks. Subjects were instructed to attend to the auditory stimulus whilst maintaining visual fixation on the speaker’s mouth for the V and AV conditions, and on a grey crosshair for the A condition. Subjects were positioned 70 cm from the screen of the monitor in a dark room and instructed to keep eye-blinks and all other motor activity to a minimum.

**EEG data acquisition and pre-processing**

EEG data were recorded at 130 locations (128 scalp + left and right mastoids) along with triggers indicating the start and end of each audio passage using an ActiveTwo system (BioSemi, Amsterdam, The Netherlands). The data were low-pass filtered online below 134 Hz and digitized at a rate of 512 Hz. Subsequent processing was conducted using MATLAB (MathWorks, Natick, USA); the data were digitally band-pass filtered between 0.5 and 25 Hz and referenced to the average of the mastoid channels. Bad channels were detected and recalculated by spline interpolating the surrounding good channels. These channels were
identified as those with a standard deviation that was greater than three times (or less than one third of) the standard deviation of the surrounding channels.

**EEG data analysis**

Two complementary methods were employed for quantifying the representation of the speech envelope in the EEG data. The first method describes the linear forward mapping from the stimulus envelope to each of the 128 scalp electrodes, known as the temporal response function (Lalor and Foxe 2010; Lalor et al. 2009). The second method describes the linear multivariate mapping from all 128 electrodes simultaneously back to the stimulus envelope, allowing reconstruction of the stimulus from unseen data (Mesgarani et al. 2009). The next two sections will describe each of these methods in further detail. Please refer to Fig. 1 for an overview of this process.

[Insert Fig. 1 about here]

**Temporal response function estimation**

Based on the assumption that the EEG response $R$ to the speech on each separate channel $n$ consisted of a convolution of the broadband speech envelope $S$ with an unknown channel-specific temporal response function (TRF) $w$, the response model can be represented in discrete time as

$$R(t,n) = \sum_{\tau} w(\tau, n) S(t - \tau) + \epsilon(t, n), \text{ for } n = 1, 2, \ldots, 128$$

(I)

where $\epsilon$ is the residual response not explained by the model. A TRF can be compared to a filter that describes the linear transformation of the continuous speech stimulus to the continuous EEG signal by the brain. Each channel-specific TRF $w(\tau, n)$ was calculated at all lags $\tau$ between $-100$ and $400$ ms relative to the stimulus onset, similar to the time interval
typically used to view an auditory evoked potential (AEP). First, the envelope $S$ was calculated using a Hilbert transform and then downsampled from 44.1 kHz to 512 Hz after applying a zero phase-shift anti-aliasing filter. Then an autocovariance matrix for the speech envelope was calculated by computing the sliding dot product of $S_t$ and $S_t^T$ across all time points and then dividing by the number of time points. A sliding dot product of $S_t$ and $R_t(n)$ was then computed and this was multiplied by the inverse of the stimulus autocovariance matrix. A regularization term was included to reduce the variance of the estimate by quadratically penalizing the difference between each two neighboring terms of $w$. This procedure can be written

\[
\begin{pmatrix}
1 & -1 \\
-1 & 2 & -1 \\
-1 & 2 & -1 \\
-1 & 2 & -1 \\
-1 & 1
\end{pmatrix}
\]

where $\langle \cdot \rangle$ indicates an average over time. The value of $\lambda$ ($4.4 \times 10^{-3}$) was empirically determined to preserve component amplitude. See (Lalor and Foxe 2010; Lalor et al. 2009) for further details.

As alluded to above, the linear assumption underlying the estimation of the TRF has implications for its interpretation. This assumption of a linear relationship between speech amplitude and EEG likely results in a response measure reflective of feedforward activity in a subset of auditory cortical cells (Lalor et al. 2009). This is in contrast to the challenge involved in disambiguating the myriad feedforward, lateral and feedback contributions to the time-locked average AEP (Di Russo et al. 2005). As such, we expect it to provide a clearer insight into the timing of activation of auditory cortex under A and AV conditions than might be possible with the AEP.
Stimulus reconstruction analysis

The relationship between the input speech dynamics and the multichannel EEG can also be captured using a multivariate analysis that offers certain advantages. Specifically, by weighting the relative contribution from each electrode simultaneously using a multivariate linear filter that incorporates the channel covariance structure in the estimation of the impulse response, one can attempt to reconstruct the speech stimulus from the EEG data using all of the available information across the scalp at each point in time. This method can result in a significant quantitative improvement in the input-output mapping relative to the TRF approach mentioned above.

To do this, a linear multivariate filter \( g \) was derived that describes an optimal mapping from the multichannel EEG data \( R \) back to the broadband speech envelope \( S \). Similar to the TRF estimation above, the filter \( g(\tau, n) \) can be calculated using normalized reverse correlation

\[
g = \left( R S^T \right)^{-1} R S^T
\]  

(3)

Given new EEG data, one can then use this filter to reconstruct an estimate of the speech stimulus that was presented during the recording of that data. In the case of \( N \) channel EEG data, this can be written as

\[
\hat{S}(t) = \sum_{n=1}^{N} \sum_{\tau} g(\tau, n) R(t - \tau, n)
\]  

(4)

where \( \hat{S} \) denotes the estimated stimulus. See (Mesgarani et al. 2009) for further details.

In this study, EEG at lags from 0 to 250 ms was used for stimulus reconstruction. The EEG and envelope were downsampled to 64 Hz to reduce processing time. For each condition, six of the seven filters were averaged and used to reconstruct the stimulus from the remaining seventh run in a leave-one-out cross-validation. The reconstruction accuracy was measured by calculating the correlation coefficient (Pearson’s \( r \)) between the reconstructed and original stimulus envelope. In order to quantify contributions from multisensory...
interactions in terms of nonlinear additivity, the AV reconstructions were compared to the sum of the unimodal (A and V) reconstructions.

RESULTS

Latency of TRF components

Unsurprisingly given the assumption underlying our TRF analysis, we observed robust TRFs for both the A and AV conditions that displayed topographic scalp distributions that were typical of auditory cortical activity. These TRFs displayed two dominant components in the A and AV condition; a negativity at around 100 ms and a positivity at around 170 ms. On all electrodes where these TRFs were apparent we observed a marked shortening in the latency of the negative component for the AV condition compared to the A condition (Fig. 2A), even at the single-subject level (Fig. 2B). Across all subjects this facilitation was of the order of 10 ms. For example, at fronto-central electrode Fz the mean latency of the early negative component was reduced by 11.4 ms in the AV condition to 98.8 ms (SD 8.9 ms) compared to 110.2 ms (SD 13.4 ms) in the A condition which was significant across all 14 subjects ($P = 2.7 \times 10^{-3}$, paired $t$-test). There was no significant shift in the later positive component ($P = 0.44$, paired $t$-test) which had mean latencies of 169.5 ms (SD 14.7 ms) for the A condition and 165.5 ms (SD 16.9 ms) for the AV condition (Fig 2A), despite the fact that such a shift was visible in those single subjects whose TRFs displayed the highest signal-to-noise ratios (Fig. 2B).

Of course, an apparent latency shift between AV and A could manifest due to additional contribution to the TRF from the visual stimulus in the AV condition. This seemed...
unlikely given the lack of any large components in the TRF corresponding to the V condition relative to both the A and AV TRFs during the timeframe of the negative component (Fig. 2A). Indeed the V TRF amplitudes in this time interval were not noticeably larger than the pre-stimulus baseline. We confirmed this at channel Fz by comparing the RMS amplitude of the V condition TRF over the pre-stimulus interval (−100 to 0 ms) with that over the interval 0 to 300 ms and finding no significant difference ($P = 0.61$, paired $t$-test). Despite this, we wished to further investigate the issue, which we did by summing the TRF measures from the A and V conditions and comparing that to the TRF from the AV condition (Fig. 3). In the A+V TRF, the mean latency of the negative component occurred at 113.6 ms (SD 10.6) which was not significantly different to that of the A condition ($P = 0.37$, paired $t$-test), but was significantly later than that of the AV condition ($P = 2.2 \times 10^{-5}$, paired $t$-test).

Importantly, this demonstrates that the facilitation we observe in the AV condition cannot simply be explained by the linear superposition of the A and V responses.

In order to further support our conclusion that what we observed was a genuine shift in the latency of the negative component, rather than an artifact caused by the addition of visual activity to the TRF in the AV condition, we examined the topographic distribution of the negative and positive components between the A and AV conditions (Fig. 4) (Delorme and Makeig 2004). Using a non-parametric randomization test (Murray et al. 2008), we found no statistical difference between these distributions for either the negative component ($P = 0.63$, TANOVA) or the positive component ($P = 0.68$, TANOVA), suggesting that they are likely to be predominantly generated by the same cortical sources. This is further illustrated by the
difference topographies (AV−A; Fig. 4) which display no signs of significant sub-additive or supra-additive activity over the frontal electrodes shown in Fig. 2A.

To examine the possibility of common generators for the negative components of the A and AV TRFs even further, we performed source analysis using the BESA software package (BESA GmbH, Gräfelfing, Germany). Specifically, we fit two symmetric dipoles to the negative components of the grand averaged A and AV TRFs and compared their locations. We based our fits on 20 ms intervals around the component peaks in each condition (A interval: 100–120 ms; AV interval: 90–110 ms). The best fitting dipoles were located at very similar locations for the A and AV conditions (A Talairach coordinates: x = ± 40.5, y = −25.8, z = 15.7; AV Talairach coordinates: x = ± 39.4, y = −19.9, z = 15). These fits explained 98.6 % and 98.9 % of the variance in the data for the A and AV TRFs, respectively. Indeed fixing the dipoles to the best-fit locations for the A TRF, and allowing their orientation to vary, still explained 98.7 % of the variance in the AV TRF, suggesting a large correspondence in the sources of both components.

Investigating nonlinear additive effects

As previously mentioned, many previous studies have interpreted nonlinear additive effects (i.e., differences between the amplitudes of AV and A+V responses) as indices of multisensory integration (Arnal et al. 2009; Besle et al. 2004; Klucharev et al. 2003; Pilling 2009; Stekelenburg and Vroomen 2007; van Wassenhove et al. 2005). Given that some of these authors have suggested that latency shifts could contribute to these observed amplitude differences (Stekelenburg and Vroomen 2007), and that we have observed a clear latency
shift when comparing the AV TRF to the summed A and V TRFs (Fig. 3), we checked for the 
existence of any amplitude differences between our response measures. Because, as 
mentioned above, the V TRF during the 0 to 300 ms interval was not significantly different 
from baseline, we opted to check for possible amplitude effects by comparing the \textit{peak} 
amplitude of both the negative and positive TRF components between the A and AV 
conditions. No difference was observed for either the negative component \((P = 0.16, \text{paired } t\text{-test})\) or the positive component \((P = 0.12, \text{paired } t\text{-test})\).

This previous analysis was based solely on the amplitude of the peak TRF components, 
which actually occurred at different latencies. Another way of relating our AV data to the 
sum of the A and V data that does not rely on a specific choice of one response feature (i.e., 
the peak), nor on choosing a specific latency, is stimulus reconstruction (Mesgarani et al. 
2009). In the present case this was done using temporally broad reconstruction filters (0–250 
ms) which captured the dominant components of both the AV and A+V conditions. As 
mentioned in the methods above, the different conditions could then be compared by 
measuring how accurately the speech envelope could be reconstructed from the 
corresponding multichannel neural response data. Reconstruction accuracy was based on the 
Pearson correlation between the reconstructed and original speech envelopes.

Figure 5 shows the distribution of single-trial reconstruction accuracy for the A, V, and 
AV conditions. In addition, for each subject and each video, we summed the corresponding A 
and V reconstructions and computed its correlation with the original speech envelope. The 
distribution of these correlation values is also shown. Considering first just the unimodal 
conditions, we found that the reconstruction accuracy of the V condition was significantly 
greater than zero \((P = 5 \times 10^{-28}, \text{one-tailed } t\text{-test})\) with a mean \(r\) value of 0.078 (SD 0.05). The 
fact that the speech envelope could be reconstructed from the V data suggests that there is a 
correlation between the dynamic changes in the visual and auditory stimuli (see next section
for further discussion). As expected, the reconstruction accuracy for the A condition was significantly higher \((P = 7.8 \times 10^{-19}, \text{paired } t\text{-test})\) with a mean \(r\) value of 0.145 (SD 0.056). In order to investigate potential nonlinear multisensory integrative processes in the AV condition, a statistical comparison was made between the A+V and AV reconstructions. Mean reconstruction accuracies of 0.161 (SD 0.049) and 0.156 (SD 0.059) were calculated for the AV and A+V conditions respectively and this difference was found not to be significant \((P = 0.34, \text{paired } t\text{-test})\). That the speech envelope could not be reconstructed more accurately in the AV condition than in the A+V condition suggests a lack of any nonlinear additive multisensory effects in our paradigm.

[Insert Fig. 5 about here]

Explaining the V reconstructions

The fact that the speech envelope reconstructions from the V data were significantly greater than zero suggests that certain characteristics of the video are correlated with the audio envelope. Based on the assumption that the contrast and luminance remained constant during each of our speech stimuli, our investigation focused on the motion within the video. While previous studies have examined the relationship between the speech envelope and mouth movements (Chandrasekaran et al. 2009), it has also been shown that even head movements are highly correlated with auditory speech (Munhall et al. 2004; Munhall and Vatikiotis-Bateson 1998). We therefore examined the frame-to-frame motion flow and its relationship with the auditory envelope for each of the seven videos used in this experiment. For each frame, a matrix of motion vectors was calculated using an Exhaustive Search (ES) block matching algorithm (Barjatya 2004). A measure of motion flow was obtained by calculating the sum of all motion vector lengths of each frame (Bartels et al. 2008). The
envelope of the auditory signal was downsampled from 44.1 kHz to 27.97 Hz to match the frame rate of the video.

Figure 6A plots the broadband speech envelope and the frame-to-frame motion flow for an 8 second segment of one of the seven videos used. This plot shows that the motion signal precedes the audio signal at the start of every sentence. A cross-correlation of these signals (Fig. 6B) across the entire duration of the video revealed a maximal correlation at a non-zero lag. This effect was found in all seven videos with the motion preceding the audio on average by ~140 ms, consistent with previously reported visual-audio lags in speech stimuli (Chandrasekaran et al. 2009; Pilling 2009; Stekelenburg and Vroomen 2007). The audio and motion signals were significantly correlated ($P < 0.05$) for six of the seven videos. If we assume that the processing of our visual motion signal is represented in the EEG (which is not necessarily true), then this correlation between auditory and visual speech (at least partially) explains how an estimate of the auditory envelope could be reconstructed from the V data with an accuracy that was significantly greater than chance. Other features of the visual stimulus that also correlated with the audio envelope and that are represented in the EEG signal may also have contributed.

While we did not see any significant TRF measure in response to the V condition (Fig. 2A), the broadly distributed lag between the visual and audio streams, peaking at ~140 ms, suggests that a V TRF may occur at some earlier time lag that was not visible in our original TRF analysis. To investigate this, we conducted an additional TRF analysis over a broader time window (~300 to 300 ms) at frontocentral electrode Fz (Fig. 6C) and occipitocentral electrode Oz (Fig. 6D). The TRFs calculated in this way did not display any large components. This is due to the fact that the TRF analysis is based on the audio envelope and the fact that the audio envelope and motion signals are correlated with each other over a broad range of non-zero lags (Fig. 6B). This weak representation of the visual motion in the
EEG signal highlights that the approach we have used is effective at acquiring response measures (i.e., the TRFs) dominated by auditory cortex in relative isolation.

DISCUSSION

We have shown a marked shortening of auditory response latency to natural continuous speech when it is accompanied by congruent visual speech. In addition, we have found no evidence of any nonlinear additive multisensory effects during natural audiovisual speech. We contend that this provides important insight into the issue of latency shifts versus amplitude effects that has been noted in previous work examining ERPs derived to discrete stimuli. In particular, the demonstrated latency shift is likely a correlate of an underlying mechanism that is fundamental to enhanced comprehension during audiovisual speech. Specifically, the fact that our latency effect is seen on the early (~100 ms) negative TRF component, a latency that has been linked with analysis of acoustic-phonetic speech features (Salmelin 2007), suggests that it may index an earlier synthesis of low level audio features into a coherent stream. This would likely allow a longer timeframe for further processing before extraction of semantic information from that stream, which has been suggested to occur after 200 ms (Salmelin 2007).

Disentangling multisensory latency and amplitude effects

Our findings fit well with a number of previous EEG/MEG studies that have proposed that visual input can speed up the processing of discrete auditory syllables (Arnal et al. 2009; Stekelenburg and Vroomen 2007; van Wassenhove et al. 2005). Specifically, these studies analyzed time-locked average ERPs/ERFs and showed that the latency of particular
component peaks was shortened for AV speech relative to A speech. It was even reported that the magnitude of this latency shortening was proportional to the degree to which the visual signal predicted the subsequent auditory stimulus (Arnal et al. 2009; van Wassenhove et al. 2005). An interesting feature of the ERPs/ERFs reported in these studies, was that the AV and A responses onset at the same latency and increased with the same slope, but that the AV responses had lower amplitude. Thus it was suggested that the peak latency shortening may simply be an artifact of this lower amplitude (Stekelenburg and Vroomen 2007). While, as mentioned above, it is difficult to disentangle putative latency effects from the putative nonlinear summation of time-locked visual and auditory responses to discrete multi-modal stimuli, our data appear to validate the existence of a latency effect in these studies. This is particularly true given that our study has utilized a data analysis approach that has produced A and AV responses that are likely dominated by auditory cortex with minimal direct contributions to the AV TRF from visual cortical activity. This is because the TRF method is based specifically on an assumption of a linear relationship between the audio envelope and the EEG. While the global motion in the video is correlated with the audio envelope, it is correlated in a temporally broad, imprecise and variable way (Fig. 6B). As such, regressing the EEG data against the envelope of the speech did not produce temporally precise TRF components in the V condition, although it still allowed reconstruction of the speech envelope from the V condition data. This implies that the sharp TRF components in the AV condition are likely to be largely generated by activity in auditory cortex. The similarities in topographic distribution and source locations between the A and AV conditions support this claim. This suggests that the differing latencies of the negative components in the A and AV conditions reflects a genuine speed up of processing in auditory cortex – a finding that is well supported by recent research in monkey auditory cortex (Chandrasekaran et al. 2013).
Previous EEG-based studies have compared the amplitude of ERPs evoked by AV stimuli to an additive model (A+V) based on the assumption that any differences may be attributed to multisensory interactions (Besle et al. 2004; Klucharev et al. 2003). What are the implications of the observed latency shift for such an approach? Each ERP component, as it is measured on the scalp, is the result of the summed activation of several neural generators, each of which may become active at different latencies (Luck 2004). For example, two separate ERPs that differ in both amplitude and latency may only appear to differ in amplitude because a certain contribution to one of the ERPs has a greater temporal overlap with activity from other sources. Thus, the interpretation of a nonlinear additive effect appears to be very difficult to validate if possible underlying latency shifts have been ignored. This is highlighted by the lack of any nonlinear additive effects in our analysis of either peak amplitude or stimulus reconstruction accuracy. It should be noted that, because we assume a linear relationship between the speech envelope and the EEG, it is likely that our TRF measures are not sensitive to nonlinear lateral and feedback contributions to speech processing. It could be that some of these contributions may show amplitude effects that are seen in the ERP. That said, the lack of any amplitude effects on the TRFs in this study is in line with recent work on audiovisual vocalizations in monkey auditory cortex that showed no firing rate amplitude effects with the addition of a visual input (Chandrasekaran et al. 2013). Also the lack of any multisensory amplitude effects may be due to the fact that our paradigm involved an easy-to-hear listening environment in which multimodal speech would be no easier to understand than unimodal speech (Ross et al. 2007a). Future work will investigate whether AV stimulus reconstructions perform significantly better than those based on the sum of A and V reconstructions for a speech-in-noise paradigm, although, interestingly, the recent monkey work also showed no multisensory amplitude effects when the SNR of the audio vocalizations was reduced (Chandrasekaran et al. 2013).
The most likely cause of the latency shift that we observe in our AV data is the precedence of visual speech; a well-known phenomenon. Stekelenburg and Vroomen (2007) showed that this effect is also present during ecologically valid non-speech stimuli if the visual stimulus preceded the auditory stimulus (e.g., handclapping), but not if it didn’t (e.g., tearing paper). There are several possible mechanisms through which the preceding visual input could cause this effect on auditory processing. One plausible process relates to the correlation between the speech envelope and visual motion. This correlation suggests that anticipatory visual motion could produce phasic variations in visual cortical activity that are relayed to auditory cortex and that correlate with the amplitude envelope of the subsequent auditory speech. This notion fits with previous work showing that the phase of oscillations in auditory cortex tracks the temporal structure of continuous visual speech (Luo et al. 2010) and fMRI data that has shown the source of the visual facilitation of speech arises from motion-sensitive cortex (Arnal et al. 2009).

Based on our EEG data, it is not possible to definitively determine the specific mechanisms or pathways by which visual activity might impact upon auditory processing. It has previously been suggested this interaction might be driven by relatively discrete visual landmarks (e.g., the onset of facial articulatory movements) that may elicit a phase-reset of ongoing low-frequency oscillations in auditory cortex, such that the arrival of the corresponding auditory syllable coincides with a high excitability phase of the auditory neuronal population (Kayser et al. 2008; Schroeder et al. 2008). The efficacy of such a mechanism in the context of continuous speech seems like it would necessitate prior knowledge about which part of an upcoming syllable was most important. For example, consider the syllable pairs bat vs vat, bat vs bet and bat vs bad. Discriminating between syllables within each of these pairs is conducted using a different part of the syllable (start,
middle and end, respectively). How a phase reset may align the high excitability phase of the auditory neuronal population with the most important part of the syllable before that syllable arrives is unclear. This is particularly true where phase-resets are occurring at a non-periodic rate in line with the frequency of the arrival of natural speech syllables (e.g., 4–7 Hz).

Another possible multisensory mechanism, which has previously been proposed in the context of the synthesis of auditory objects, is that the particular neural populations may be sensitive to temporal correlations between the speech envelope and visual motion leading to them becoming integrated into a coherent multisensory speech object. Recent work has shown that cortical entrainment to the speech envelope is not just representative of bottom-up envelope tracking, but that it may index an analysis-by-synthesis approach to speech processing (Ding et al. 2013). This framework proposes that speech is first analyzed by breaking it up into constituent spectrotemporal channels and that auditory objects are synthesized from those channels that modulate together in a temporally coherent manner. Within such a framework, Ding et al., (2013) have shown that cortical entrainment to speech depends on the detailed spectro-temporal fine structure of speech and that, as such, the cortical entrainment to the speech envelope may represent the temporally coherent fluctuations of the many spectro-temporal features that make up the speech signal. This fits with a theory postulating that in auditory scene analysis, multi-feature auditory sources are segregated into perceptual streams based on the temporal coherence of such acoustic features (Shamma et al. 2011). In keeping with previous work espousing a correlated mode of processing for audiovisual speech (Campbell 2008), we postulate that motion in the visual stream, being correlated with the dynamics of the auditory features, results in the visual signal being bound to the audio features to form a multisensory object. In terms of intelligibility, this partially explains the benefit conferred by the presence of visual speech in adverse acoustic conditions. The fact that the visual motion precedes the auditory features
may mean that the binding of auditory features occurs earlier and that this synthesis is reflected in a shorter latency of our envelope tracking measure. Such a suggestion also seems to be compatible with the idea of predictive coding (Arnal et al. 2009; van Wassenhove et al. 2005). Of course the TRF we observe is still derived from the envelope of the heard speech. Thus, the latency shift we obtain (~10 ms) is but a small fraction of the actual amount by which the video preceded the audio (>100 ms). This discrepancy in timings is also likely to be a result of the time it takes for motion information from the visual cortex to have an impact upon auditory processing.

In terms of what specific neural populations might facilitate the binding of temporally coherent visual and audio speech, one candidate region would the superior temporal sulcus (STS), which has previously been linked with multisensory speech processes (Beauchamp et al. 2004). Indeed recent research has shown evidence for neural computations in this area that underpin auditory figure-ground segregation using stimuli that display periods of temporal coherence across multiple frequency channels (Teki et al. 2011). Of course it remains a possibility that audiovisual speech interactions could occur through direct connections (Schroeder et al. 2008). The latency of our TRF effects (~100 ms) suggests that this may indeed be true (Arnal et al. 2009). Further research using alternative brain recording methods will elucidate these issues.

Comparisons with other findings and future applications

Our findings are somewhat in contrast with recent MEG work that also used continuous auditory speech to examine the effects of visual input on TRFs acquired to A and AV speech (Zion-Golumbic et al. 2013). This study reported no latency effects on TRFs that were dominated by a single peak at around 50 ms, which is considerably earlier than the components that we have focused on here. This difference may simply be a result of the
sensitivities of EEG and MEG to different processes within the auditory system. A possible explanation for the lack of a latency effect on such an early component is that it is so early that it may only index the analysis phase of the analysis-by-synthesis process, i.e., the phase wherein the basic acoustic features of the auditory signal are extracted subcortically (Nelken 2008). Processing at this early stage may be unaffected by the dynamics of a simultaneously presented visual input.

While it will certainly require methods complementary to human EEG to determine the details of the neural mechanisms underlying audiovisual speech integration, we suggest that the latency effect that we have identified may index these underlying mechanisms. This robust index, acquired using natural stimuli and a relatively inexpensive and accessible technology, may prove useful in research with clinical populations in which altered multisensory (AV) processing has been reported, e.g., dyslexia (Hairston et al. 2005), autism (Brandwein et al. 2012), and schizophrenia (Ross et al. 2007b).

ACKNOWLEDGEMENTS

The authors would like to thank James O’Sullivan, Gerard Loughnane and Fiona Newell for critical discussion and reading of the manuscript. This work was supported by the Irish Higher Education Authority’s Graduate Research Education Programme in Engineering and by a grant from Science Foundation Ireland (09-RFP-NES2382).

REFERENCES


Millman RE, Prendergast G, Hymers M, and Green GGR. Representations of the temporal envelope of sounds in human auditory cortex: Can the results from invasive


FIGURE CAPTIONS

FIG. 1. Linear forward and backward mapping techniques. The temporal response function (TRF), which represents the mapping from the acoustic envelope to the EEG data at each channel, was calculated using a linear least squares approach. The latencies of defined TRF
components were measured and compared between A and AV speech. All 128 channels of EEG were used simultaneously to reconstruct an estimate of the acoustic envelope from unseen EEG data using a reverse mapping technique. Reconstruction accuracy was measured by calculating the Pearson’s correlation between the reconstructed estimate and the original envelope.

FIG. 2. Comparison of bimodal and unimodal TRFs. A: Grand average TRFs calculated at several frontocentral sites (C3, FC3, Fz, FC4 and C4) and at occipital site (Oz) for V, A and AV speech. At each of the frontocentral sites, the negative component of the AV response (blue trace) is noticeably earlier than that of the A response (green trace) at each of the frontocentral sites, while the TRFs calculated from the V data (orange trace) are considerably reduced in amplitude. No significant TRF measure was calculated at occipital location Oz for any of the 3 conditions. B: TRF measures calculated at Fz for two subjects who displayed the highest signal-to-noise ratios. The latency shift is still clearly visible at the single subject level and is even evident in the later positive component.

FIG. 3. Comparison of bimodal and summed unimodal TRFs. Grand average TRFs calculated at the same frontocentral sites (C3, FC3, Fz, FC4 and C4) and occipital site Oz for AV speech and for a summation of the A and V responses (A+V). The negative component of the AV response (blue trace) remains earlier than that of the A+V response (brown trace). Again, no significant TRF measure was calculated at occipital location Oz for either of the 2 conditions.

FIG. 4. Topographical distributions of dominant TRF components for A and AV speech. The distribution of the negative component for A (~110 ms) and AV (~100 ms) speech is shown in the top left and top middle panels respectively, while the distribution of the positive
component for A (~170 ms) and AV (~165 ms) speech is shown in the bottom left and bottom middle panels respectively. Difference topographies (AV−A) are shown for the negative component (top right) and positive components (bottom right). The same amplitude scale is used for all topographic maps and is indicated by the color bars.

FIG. 5. Investigating nonlinear additivity to quantify multisensory interactions. Distribution of reconstruction accuracy (Pearson’s $r$) for the V, A, A+V and AV condition. Each A+V measure was obtained by summing the paired reconstructions from the A and V conditions and calculating the correlation between the resulting estimate and the original envelope. The dashed line corresponds to the mean value ($V = 0.078; A = 0.145; A+V = 0.156; AV = 0.161$).

FIG. 6. Temporal lag between visual and auditory speech. A: Auditory speech envelope (grey trace) and frame-to-frame motion flow (black trace) for an 8 second time segment of one of the videos used in the experiment. There is a clear correspondence between the auditory envelope and the motion flow. The temporal lag is also very evident at the start of each sentence. B: Cross-correlation of the same envelope and motion signals. The dotted line represents a zero time lag relative to the motion signal. C: Grand average TRFs calculated at frontocentral site Fz for an extended time window (−300 to 300 ms). There is no evidence of visual-evoked TRFs at these earlier lags. D: Grand average TRFs calculated at occipitocentral site Oz over the same time window. Again there is no evidence, even at this occipital site, of any visual-evoked TRFs.