Systems/Circuits

Congruent Visual Speech Enhances Cortical Entrainment to Continuous Auditory Speech in Noise-Free Conditions

Michael J. Crosse,[1,2] John S. Butler,[1,2,4] and Edmund C. Lalor[1,2,3]

[1]School of Engineering, [2]Trinity Centre for Bioengineering, and [3]Trinity College Institute of Neuroscience, Trinity College Dublin, Dublin 2, Ireland, and [4]School of Mathematical Sciences, Dublin Institute of Technology, Dublin 8, Ireland

Congruent audiovisual speech enhances our ability to comprehend a speaker, even in noise-free conditions. When incongruent auditory and visual information is presented concurrently, it can hinder a listener's perception and even cause him or her to perceive information that was not presented in either modality. Efforts to investigate the neural basis of these effects have often focused on the special case of discrete audiovisual syllables that are spatially and temporally congruent, with less work done on the case of natural, continuous speech. Recent electrophysiological studies have demonstrated that cortical response measures to continuous auditory speech can be easily obtained using multivariate analysis methods. Here, we apply such methods to the case of audiovisual speech and, importantly, present a novel framework for indexing multisensory integration in the context of continuous speech. Specifically, we examine how the temporal and contextual congruency of ongoing audiovisual speech affects the cortical encoding of the speech envelope in humans using electroencephalography. We demonstrate that the cortical representation of the speech envelope is enhanced by the presentation of congruent audiovisual speech in noise-free conditions. Furthermore, we show that this is likely attributable to the contribution of neural generators that are not particularly active during unimodal stimulation and that it is most prominent at the temporal scale corresponding to syllabic rate (2–6 Hz). Finally, our data suggest that neural entrainment to the speech envelope is inhibited when the auditory and visual streams are incongruent both temporally and contextually.

Key words: **audiovisual speech; EEG; multisensory integration; stimulus reconstruction; temporal coherence; temporal response function**

Significance Statement

Seeing a speaker's face as he or she talks can greatly help in understanding what the speaker is saying. This is because the speaker's facial movements relay information about what the speaker is saying, but also, importantly, when the speaker is saying it. Studying how the brain uses this timing relationship to combine information from continuous auditory and visual speech has traditionally been methodologically difficult. Here we introduce a new approach for doing this using relatively inexpensive and noninvasive scalp recordings. Specifically, we show that the brain's representation of auditory speech is enhanced when the accompanying visual speech signal shares the same timing. Furthermore, we show that this enhancement is most pronounced at a time scale that corresponds to mean syllable length.

Introduction

During natural, everyday conversation, we routinely process speech using both our auditory and visual systems. The benefit of viewing a speaker's articulatory movements for speech comprehension has been well documented and has been characterized in terms of two specific modes of audiovisual (AV) information: "complementary" and "correlated" (Summerfield, 1987; Campbell, 2008). Visual speech assumes a complementary role when it is required to compensate for underspecified auditory speech, enhancing perception, e.g., in adverse hearing conditions (Sumby and Pollack, 1954; Ross et al., 2007a) and in people with impaired hearing (Grant et al., 1998). It assumes a correlated role when there is redundancy between the information provided by vision and audition, e.g., in optimal listening conditions where it has been shown to benefit people with normal hearing (Reisberg et al., 1987). Specifically, in the latter case, enhanced perception is

Received May 12, 2015; revised Aug. 12, 2015; accepted Sept. 8, 2015.

Author contributions: M.J.C. and E.C.L. designed research; M.J.C. performed research; M.J.C. analyzed data; M.J.C., J.S.B., and E.C.L. wrote the paper.

This work was supported by the Programme for Research in Third-Level Institutions and cofunded under the European Regional Development fund. We thank Shane Hunt for assisting in the design and manufacture of experimental hardware.

The authors declare no competing financial interests.

Correspondence should be addressed to Dr. Edmund C. Lalor, Trinity College Dublin, Dublin 2, Ireland. E-mail: edlalor@tcd.ie.

DOI:10.1523/JNEUROSCI.1829-15.2015

 0270-6474/15/3514195-10$15.00/0

possible because the visible articulators that determine the vocal resonances, such as the lips, teeth, and tongue, as well as ancillary movements, such as facial, head, and hand movements, are temporally correlated with the vocalized acoustic signal (Summerfield, 1992; Grant and Seitz, 2000; Jiang and Bernstein, 2011). However, relatively little research has explicitly examined how the temporal correlation between auditory and visual speech affects the neural processing of continuous AV speech.

Electroencephalography (EEG) and magnetoencephalography (MEG) studies have demonstrated that auditory cortical activity entrains to the temporal envelope of speech (Ahissar et al., 2001; Abrams et al., 2008; Lalor and Foxe, 2010). Although many studies have examined the effects of attention on envelope tracking (Ding and Simon, 2012; Power et al., 2012; Sheedy et al., 2014), less work has examined how this process may be influenced by visual speech [but see Zion Golumbic et al. (2013) and Luo et al. (2010)]. Traditionally, EEG/MEG studies have focused on how the brain responds to discrete AV stimuli such as syllables (Sams et al., 1991; Möttönen et al., 2002), an approach that is limited in what it can say about the role of the temporal correlation between continuous auditory and visual speech. Indeed, many EEG/MEG studies have reported interesting cross-modal interaction effects on cortical response measures, even when the discrete stimuli were phonetically incongruent (Klucharev et al., 2003; van Wassenhove et al., 2005; Stekelenburg and Vroomen, 2007; Arnal et al., 2009). This is unsurprising, given that particular combinations of incongruent AV syllables elicit illusory percepts when presented concurrently (McGurk and MacDonald, 1976). It has been suggested (Campbell, 2008) that because such discrete incongruent stimuli are spatially and temporally coherent and coextensive, this may act as a cue to their integration.

Here, we used natural, continuous speech stimuli, allowing us to examine how EEG entrains to temporally and contextually congruent and incongruent AV speech. Specifically, we hypothesize that the benefits of congruent AV speech will be detectable in noise-free conditions and indexed by enhanced envelope tracking. We also implement several follow-up experiments to answer the following research questions: (1) Is a dynamic human face sufficient to enhance envelope tracking, even when it is temporally incongruent? (2) Does contextually incongruent information, such as conflicting gender, modulate envelope tracking differently? (3) Is any dynamic visual stimulus sufficient to enhance envelope tracking, even if it does not comprise a human face? (4) Conversely, does a static human face enhance the tracking of a dynamic auditory input?

To obtain a direct measure of envelope tracking, we reconstructed an estimate of the speech envelope from the EEG data and compared it with the original envelope (Rieke et al., 1995; Mesgarani et al., 2009). One of the main goals of this study was to establish a framework for quantifying multisensory interactions using this stimulus reconstruction approach. Within this framework, we also investigated how our measures of multisensory interaction varied across different temporal scales with a view to elucidating whether the effects were more prominent at any particular level of speech processing (i.e., phonemic, syllabic, word, prosodic; Giraud and Poeppel, 2012).

Materials and Methods

Participants. Twenty-one native English speakers (eight females; age range, 19–37 years) participated in the experiment. Written informed consent was obtained from each participant beforehand. All participants were right-handed, were free of neurological diseases, had self-reported normal hearing, and had 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.

Table 1. Experimental conditions and stimulus content

	Stimuli	
Condition	Audio	Video
A	Male speaker	Black screen with gray fixation crosshair
V	None	Male speaker
AVc	Male speaker	Congruent male speaker
AVi	Male speaker	Incongruent male speaker
AVif	Male speaker	Incongruent female speaker[a]
AVin	Male speaker	Wildlife scenes with fixation crosshair
AVsf	Male speaker	Still image of male speaker's face

[a]A different female speaker was used in each of the 15 trials to prevent association with the male speaker's voice.

Stimuli and procedure. The speech stimuli were drawn from a collection of videos featuring a trained male speaker. The videos consisted of the speaker's head, shoulders, and chest, centered in the frame. Speech was directed at the camera, and the speaker used frequent, but natural, hand movements. There was no background movement or noise. The speech was conversational-like and continuous, with no prolonged pauses between sentences. The linguistic content centered on political policy, and the language was colloquial American English. Fifteen 60 s videos were rendered into 1280 × 720 pixel movies in VideoPad Video Editor (NCH Software). Each video had a frame rate of 30 frames per second, and the soundtracks were sampled at 48 kHz with 16-bit resolution. Dynamic range compression was applied to each soundtrack in Audacity audio editor such that lower intensities of the speech signal could be amplified. Compression was applied at a ratio of 10:1 above a threshold of −60 dB. The signal was only amplified above a noise floor of −45 dB, which prevented the gain increasing during pauses and unduly amplifying breathing sounds. The intensity of each soundtrack, measured by root mean square, was normalized in MATLAB (MathWorks).

To test the main hypothesis of the study and the four follow-up questions posed in the Introduction, we dubbed the same 15 soundtracks to five different kinds of visual stimuli. (1) Congruent audiovisual stimuli (AVc) were created by redubbing each soundtrack to its original video, i.e., A1V1, A2V2, etc. Unimodal versions were also produced as a control, i.e., audio-only stimuli (A) and visual-only stimuli (V). (2) To examine the role of temporal congruency, incongruent audiovisual stimuli (AVi) were created by mismatching the same 15 soundtracks and videos, i.e., A1V2, A2V3, etc. (3) To examine the role of contextual congruency, the soundtracks were dubbed to videos of incongruent female speakers (AVif). The female speakers were centered in the frame (head, shoulders, and chest), and their speech was directed at the camera. (4) To examine the impact of a dynamic (nonhuman) visual stimulus, incongruent nature stimuli (AVin) were created by dubbing the speech soundtracks to wildlife documentaries. (5) To examine the role of human-specific visual features, the soundtracks were dubbed to still images of the male speaker's static face (AVsf). For a summary of all the stimuli used in the experiment, please refer to Table 1.

Stimulus presentation and data recording took place in a dark sound-attenuated room with participants seated at a distance of 70 cm from the visual display. Visual stimuli were presented on a 19 inch CRT monitor operating at a refresh rate of 60 Hz. Audio stimuli were presented diotically through Sennheiser HD650 headphones at a comfortable level of ~65 dB. Stimulus presentation was controlled using Presentation software (Neurobehavioral Systems). Each of the 15 speech passages was presented seven times, each time as part of a different experimental condition (Table 1). Presentation order was randomized across conditions, within participants. Participants were instructed to fixate on either the speaker's mouth (V, AVc, AVi, AVif, AVsf) or a gray crosshair (A, AVin) and to minimize eye blinking and all other motor activity during recording.

To encourage active engagement with the content of the speech, participants were required to respond to target words via button press. Before each trial, a target word was displayed on the monitor until the

participant was ready to begin. All target words were detectable in the auditory modality except during the V condition, where they were only visually detectable. Hits were counted for responses that were made 200–2000 ms after the onset of auditory voicing, and feedback was given at the end of each trial. A target word could occur between one and three times in a given 60 s trial, and there were exactly 30 targets in total per condition. A different set of target words was used for each condition to avoid familiarity, and assignment of target words to the seven conditions was counterbalanced across participants.

Behavioral data analysis. Participants' performance on the target detection task was examined for multisensory effects. Specifically, we examined whether reaction times (RTs) were facilitated by congruent bimodal speech (AVc) compared with unimodal speech (A, V), an effect known as a redundant signals effect (RSE). An RSE does not necessarily imply multisensory interaction unless it violates the race model (Raab, 1962). The race model predicts that the RT in response to a bimodal stimulus is determined by the faster of the two unimodal processes. Violation of the race model was examined using the following inequality (Miller, 1982):

$$F_{\mathrm{AVc}}(t) \leq F_{\mathrm{A}}(t) + F_{\mathrm{V}}(t), \quad t > 0, \tag{1}$$

where F_{AVc}, F_{A}, and F_{V} are the cumulative distribution functions (CDFs) based on the RTs of the AVc, A, and V conditions, respectively. CDFs were generated for each participant and condition, divided into nine quantiles (0.1, 0.2, . . . , 0.9) and group averaged (Ulrich et al., 2007).

EEG acquisition and preprocessing. Continuous EEG data were acquired using an ActiveTwo system (BioSemi) from 128 scalp electrodes and two mastoid electrodes. The data were low-pass filtered on-line below 134 Hz and digitized at a rate of 512 Hz. Triggers indicating the start of each trial were recorded along with the EEG. These triggers were sent by an Arduino Uno microcontroller, which detected an audio click at the start of each soundtrack by sampling the headphone output from the PC. Subsequent preprocessing was conducted off-line in MATLAB; the data were bandpass filtered between 0.3 and 30 Hz, downsampled to 64 Hz, and rereferenced to the average of the mastoid channels. To identify channels with excessive noise, the time series were visually inspected in Cartool (brainmapping.unige.ch/cartool), and the SD of each channel was compared with that of the surrounding channels in MATLAB. Channels contaminated by noise were recalculated by spline-interpolating the surrounding clean channels in EEGLAB (Delorme and Makeig, 2004).

Because our aim was to examine how visual information affects the neural tracking of auditory speech, the stimuli were characterized using the broadband envelope of the acoustic signal (Rosen, 1992). To model the input to the auditory system, the stimuli were first bandpass filtered into 128 logarithmically-spaced frequency bands between 100 and 6500 Hz using a gammatone filterbank. The uppermost and lowermost filter limits captured the first, second, and third formant spectral regions of the speech signals, known to carry the acoustic information that correlates most with visual speech features (Grant and Seitz, 2000; Chandrasekaran et al., 2009). The envelope at each of the 128 frequency bands was calculated using a Hilbert transform, and the broadband envelope was obtained by averaging over the 128 narrowband envelopes.

Stimulus reconstruction. To determine how faithfully the cortical activity tracked the speech envelope during each condition, we measured the accuracy with which we could reconstruct the envelope from the EEG data. Suppose the EEG response at electrode n and at time $t = 1 \ldots T$ is represented as $r_n(t)$ and the stimulus envelope as $s(t)$. The reconstruction filter, $g_n(\tau)$, represents the linear mapping from $r_n(t + \tau)$ to $s(t)$ at time lag τ and can be expressed as follows:

$$\hat{s}(t) = \sum_n \sum_\tau r_n(t + \tau)\, g_n(\tau), \tag{2}$$

where $\hat{s}(t)$ is the estimated stimulus envelope. Here, the entire filter, $\mathbf{g}$, was obtained for all 128 electrodes simultaneously using ridge regression, written in matrix form as follows:

$$\mathbf{g} = (\mathbf{R}^{\mathrm{T}}\mathbf{R} + \lambda\mathbf{I})^{-1}\,\mathbf{R}^{\mathrm{T}}\mathbf{s}, \tag{3}$$

where $\mathbf{R}$ is the lagged time series of the EEG data and can be defined as follows:

$$\mathbf{R} = \begin{bmatrix} r_1(\tau_{\max}+1) & \ldots & r_{128}(\tau_{\max}+1) & r_1(\tau_{\max}) & \ldots & r_{128}(\tau_{\max}) & \ldots & r_1(1) & \ldots & r_{128}(1) \\ \vdots & & \vdots & \vdots & & \vdots & & \vdots & & \vdots \\ r_1(T) & \ldots & r_{128}(T) & r_1(T-1) & \ldots & r_{128}(T-1) & \ldots & r_1(T-\tau_{\max}) & \ldots & r_{128}(T-\tau_{\max}) \\ 0 & \ldots & 0 & r_1(T) & \ldots & r_{128}(T) & \ldots & r_1(T-\tau_{\max}+1) & \ldots & r_{128}(T-\tau_{\max}+1) \\ 0 & \ldots & 0 & 0 & \ldots & 0 & \ldots & r_1(T-\tau_{\max}+2) & \ldots & r_{128}(T-\tau_{\max}+2) \\ \vdots & & \vdots & \vdots & & \vdots & & \vdots & & \vdots \\ 0 & \ldots & 0 & 0 & \ldots & 0 & \ldots & r_1(T) & & r_{128}(T) \end{bmatrix}. \tag{4}$$

The time lags τ ranged from 0 to 500 ms poststimulus, i.e., $\tau_{\max} = 32$ samples. A constant term was included in the regression model by concatenating 128 columns of ones to the left of $\mathbf{R}$. The regularization term in Eq. 3 was used to prevent overfitting to noise along the low-variance dimensions where λ was the ridge parameter and $\mathbf{I}$ was the identity matrix.

The regression analysis was performed using a custom-built toolbox in MATLAB (mTRF Toolbox, version 1.2; http://www.mee.tcd.ie/lalorlab/resources.html). Leave-one-out cross-validation was used to reconstruct an estimate of each of the 15 stimuli per condition. Reconstruction accuracy was measured by calculating the correlation coefficient between the estimated and original speech envelopes. To optimize performance within each condition, we conducted a parameter search (over the range $2^{14}, 2^{15}, \ldots, 2^{21}$) for the λ value that maximized the correlation between $\hat{s}(t)$ and $s(t)$. To prevent overfitting, λ was tuned to the value that gave the highest mean reconstruction accuracy across the 15 trials.

Quantifying multisensory interactions. Our decision to include all 128 channels of EEG in the reconstruction analysis is justified because irrelevant filter channels can maintain zero weight while allowing the model to capture additional variance (Pasley et al., 2012). However, this multichannel approach required us to apply different criteria when quantifying multisensory interactions in the congruent and incongruent AV conditions. For the incongruent AV conditions (AVi, AVif, AVin, AVsf), a maximum model criterion was applied, i.e., each multisensory condition was compared with the optimal unisensory (A) condition. This was fair because the incongruent visual stimuli were not temporally correlated with the speech envelope; therefore, information encoded by the visual system in occipital channels did not benefit reconstruction of the envelope. However, this was not true for the congruent AV condition (AVc), where the dynamics of the visual stimulus were highly correlated with those of the speech envelope. This would allow the AVc model to infer complementary information from correlated visual speech processing as reflected on parieto-occipital channels (Luo et al., 2010; Bernstein and Liebenthal, 2014), even without ever explicitly quantifying those visual features in the model fitting. Previous work has attempted to circumvent this bias by restricting the analysis to only the frontal electrodes (O'Sullivan et al., 2013). However, this approach significantly compounds model performance and, in any case, would not guarantee that the AVc condition was unbiased as volume conduction could still result in visual cortical activity being reflected in frontal channels.

Instead, we examined multisensory interactions in the AVc condition using the additive model criterion (Stein and Meredith, 1993). The rationale here is that multisensory interactions can be inferred from differences between cortical responses to multisensory stimuli and the

summation of unisensory responses [i.e., AVc − (A + V)]. The validity of the additive model for the purpose of indexing multisensory integration in electrophysiological studies is well established (Besle et al., 2004a). The following procedure was used to apply the additive model approach to our stimulus reconstruction analysis. (1) New A and V reconstruction filters were calculated using the A and V data sets, respectively ($\lambda_A = 2^{14}, 2^{15},\ldots, 2^{20}$; $\lambda_V = 2^{14}, 2^{15},\ldots, 2^{34}$). (2) We calculated the algebraic sum of the A and V filters (A+V) for every combination of λ values. (3) Critically, each additive model was then assessed using the EEG data from the AVc condition; this ensured that the model could decode the envelope from channels that encoded both auditory and visual information. (4) A grid search was conducted to find the combination of λ values that maximized reconstruction accuracy across the 15 stimuli. The difference between the AVc and A+V models was quantified in terms of how accurately each of them could reconstruct the speech envelopes from the AVc data using leave-one-out cross-validation. We interpreted such differences as an index of multisensory interaction.

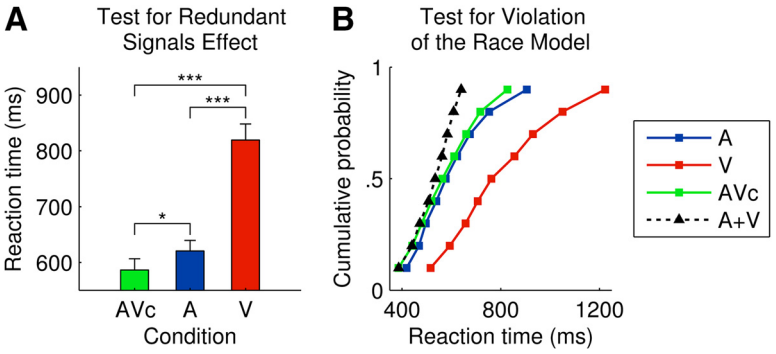

Figure 1. Examination of behavior under the race model. ***A***, Mean ($N = 21$) reaction times for the AVc (green), A (blue), and V (red) conditions. Error bars indicate SEM across participants. Brackets indicate pairwise statistical comparisons (*$p < 0.05$; ***$p < 0.001$). ***B***, Group-average ($N = 18$) cumulative distribution functions based on the reaction times shown in ***A***. The dashed black trace represents the facilitation predicted by the race model (A+V).

Temporal response function estimation. To visualize the temporal profile of the neural response to the different stimuli, we calculated the temporal response function (TRF) at every channel. A TRF can be interpreted as a filter, **w**, that describes the brain's linear transformation of the speech envelope to the continuous neural response at each channel location. Unlike the stimulus reconstruction approach, it is not a multivariate regression but represents multiple univariate mappings between stimulus and EEG. TRF model parameters are neurophysiologically interpretable, i.e., nonzero weights are only observed at channels where cortical activity is related to stimulus encoding (Haufe et al., 2014). This allows for examination of the amplitude, latency, and scalp topography of the stimulus–EEG relationship, complementing the stimulus reconstruction approach. For each 60 s trial, the TRFs were calculated at time lags between −100 and 400 ms as follows:

$$\mathbf{w} = (\mathbf{S}^{\mathrm{T}}\mathbf{S} + \lambda\mathbf{M})^{-1}\,\mathbf{S}^{\mathrm{T}}\mathbf{r}, \qquad (5)$$

where **S** is the lagged time series of the stimulus envelope and **r** is the EEG response data. The regularization term, **M**, used to prevent overfitting also preserved component amplitude by penalizing the quadratic term. The ridge parameter, λ, was empirically chosen to maintain component amplitude [see Lalor and Foxe (2010) for further details].

Multidimensional scaling. In an effort to visualize any potentially interpretable differences between the various reconstruction models, we applied nonmetric multidimensional scaling (MDS) to the model channel weights. MDS has been applied to electrophysiological data in previous studies to demonstrate the dissimilarity of neural responses elicited to different phonemes (Chang et al., 2010; Di Liberto et al., 2015). Given a set of objects, MDS works by embedding each object in a multidimensional space such that distances between objects produce an empirical matrix of dissimilarities. Here, the objects are the different stimulus conditions, and the dissimilarities are the standardized Euclidean distances between the filter weights. To capture maximal model variance across the scalp, weight vectors from all 128 channels were concatenated and group averaged. To determine how many dimensions would be maximally required to explain model variance, Kruskal's stress was measured as a function of dimensions (Kruskal and Wish, 1978). Two dimensions were sufficient to meet the criteria, i.e., stress <0.1.

Statistical analyses. Any effects of condition on behavior or EEG measures were established using one-way repeated-measures ANOVAs, except where otherwise stated. Where sphericity was violated, the Greenhouse–Geisser-corrected degrees of freedom are reported. *Post hoc* comparisons were conducted using two-tailed (paired) *t* tests, except where one-tailed tests were necessary. Multiple pairwise comparisons were corrected for using the Holm–Bonferroni method. All numerical values are reported as mean ± SD.

Results

Behavior

Twenty-one participants performed a target detection task during EEG recording. To examine whether the detection of auditory targets was affected by the visual stimulus, we compared the reaction times and hit rates across the five AV conditions (AVc, AVi, AVif, AVin, AVsf). The visual stimulus had a significant effect on RT ($F_{(4,80)} = 3.13$, $p = 0.02$) but not on hit rate, which was near ceiling (median, >92%; $\chi^2_{(4)} = 7.49$, $p = 0.11$, Friedman test). To test for an RSE, planned *post hoc* comparisons were made between the congruent AV condition (AVc) and the unimodal conditions (A, V; Fig. 1*A*). RTs for the AVc condition (586 ± 92 ms) were significantly faster than those for both the A condition (620 ± 88 ms; $t_{(20)} = 2.74$, $p = 0.01$) and the V condition (819 ± 136 ms; $t_{(20)} = 7.9$, $p = 1.4 \times 10^{-7}$), confirming an RSE. To test whether this RSE exceeded the statistical facilitation predicted by the race model, we compared the bimodal CDFs with the sum of the unimodal CDFs (Fig. 1*B*). Three participants were excluded from this analysis as they did not detect enough targets in the V condition to allow estimation of the CDF. The race model was violated by >50% of participants at the first two quantiles, but the effect was not significant (first quantile: $t_{(17)} = 0.01$, $p = 0.5$; second quantile: $t_{(17)} = 0.16$, $p = 0.56$; one-tailed tests). This is likely attributable to the nature of our task involving, as it did, an easy auditory detection task and much more difficult visual detection (lipreading) task. As such, RTs in the AVc condition were likely dominated by reaction to the auditory stimulus with minimal contribution from the visual modality. None of the incongruent AV conditions (AVi, AVif, AVin, AVsf) showed behavioral differences relative to the A condition or each other.

Impact of AV congruency on envelope tracking

To investigate the impact of AV congruency on the cortical representation of speech, we reconstructed an estimate of the speech envelope from the EEG data for each condition (Fig. 2*A*). Critically, we found that the envelope was encoded more accurately by congruent AV speech (AVc; Pearson's $r = 0.2 \pm 0.05$) than could be explained by our additive model (A+V; 0.18 ± 0.04; $t_{(20)} = 3.84$, $p = 0.001$; Fig. 2*B*). This suggests that, even in optimal listening conditions, congruent visual speech enhances neural tracking of the acoustic envelope in line with our primary hypothesis.

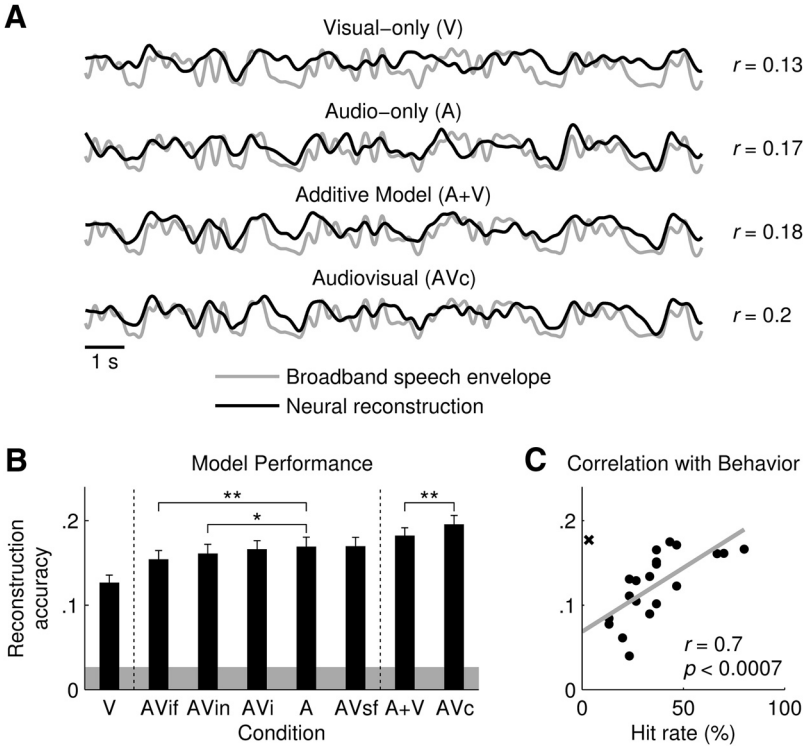

Figure 2. Reconstruction of the speech envelope from EEG. ***A***, Examples of the original speech envelope (gray) with the group-average neural reconstruction (black) superimposed. Signals were filtered below 3 Hz for visualization. The mean correlation coefficient between the original and reconstructed envelopes (i.e., reconstruction accuracy) is shown to the right. ***B***, Mean ($N = 21$) reconstruction accuracy for all eight models in ascending order. Error bars indicate SEM across participants. Dashed lines indicate planned *post hoc* subgroups, and brackets indicate pairwise statistical comparisons (*$p < 0.05$; **$p < 0.01$). The shaded area represents the 95th percentile of chance-level reconstruction accuracy (permutation test). ***C***, Correlation ($N = 20$) between reconstruction accuracy and hit rate using visual speech data. Each data point represents a participant's mean value, and the × marker indicates the participant that was excluded from the analysis. The gray line represents a linear fit to the data.

As discussed above, quantifying multisensory interactions in the incongruent AV conditions (AVi, AVif, AVin, AVsf) simply involved direct comparisons with the A condition. Across these five conditions, there was a significant effect of visual stimulus on reconstruction accuracy ($F_{(2,40.3)} = 11.84$, $p = 8.8 \times 10^{-5}$; Fig. 2*B*). However, *post hoc* comparisons revealed that envelope tracking was not enhanced by incongruent AV speech relative to unimodal speech. This suggests that the neural mechanism underlying enhanced envelope tracking in the case of congruent AV speech relies on discrete, phasic interactions as opposed to an ongoing, tonic process; in other words, it is likely that the temporal coherence between auditory and visual speech is of paramount importance to this multisensory enhancement. Although we did not find an enhancement effect, we did find that envelope tracking was actually inhibited by incongruent AV speech, but only when the visual stimulus was incongruent both temporally and contextually. Relative to the A condition (0.17 ± 0.05), envelope tracking was significantly inhibited by the presentation of an incongruent female speaker (AVif; 0.15 ± 0.05; $t_{(20)} = 3.3$, $p = 0.004$) and incongruent nature scenes (AVin; 0.16 ± 0.05; $t_{(20)} = 2.3$, $p = 0.03$). Unsurprisingly, envelope tracking was worst in the V condition (0.13 ± 0.04); however, it maintained accuracy significantly above the 95th percentile of chance level (Fig. 2*B*, shaded area). This demonstrates the efficacy of the stimulus reconstruction method to infer temporally correlated information pertaining to one sensory modality from another.

Recently, Ding and Simon (2013) demonstrated that the accuracy with which the envelope can be reconstructed from MEG data is highly correlated with stimulus intelligibility across participants. This could only be demonstrated at a signal-to-noise ratio (SNR) where intelligibility scores were at an intermediate level, i.e., ~50%. In our study, the V condition was the only one where hit rate was not at ceiling ($36.8 \pm 18.1\%$). Under the assumption that hit rate is also reflective of intelligibility, we calculated the correlation coefficient between each participant's mean reconstruction accuracy and hit rate using the V data (Fig. 2*C*). We found that this measure of behavior was significantly correlated with reconstruction accuracy across participants ($r = 0.7$, $p = 6.6 \times 10^{-4}$). Participant 13 was excluded from this analysis as an outlier as the participant reported an inability to detect any targets during the V condition (Fig. 2*C*, × marker).

Multisensory interaction effects as a function of temporal scale

It has been suggested that AV speech perception includes the neuronal integration of temporally fine-grained correlations between the auditory and visual speech stimuli, even at the phonetic level (Grant and Seitz, 2000; Klucharev et al., 2003). In contrast, other work has suggested that, at least in some detection paradigms, neuronal integration at this detailed level of granularity is not necessary (Tjan et al., 2014). We attempted to elucidate whether our multisensory effects [i.e., AVc > (A + V)] may be occurring on the time scale of phonemes, syllables, words, or sentences. To do this, we calculated the correlation coefficient between the reconstructed and original envelopes at every 2-Hz-wide frequency band between 0 and 30 Hz. Figure 3*A* shows reconstruction accuracy as a function of frequency for the AVc and A+V models, whereas Figure 3*B* shows the multisensory interaction effect [AV − (A + V)] at each frequency band. Significant multisensory effects were measured at 2–4 Hz ($t_{(20)} = 4.74$, $p = 1.3 \times 10^{-4}$) and 4–6 Hz ($t_{(20)} = 4.1$, $p = 5.6 \times 10^{-4}$). This suggests that neural tracking of the acoustic envelope is enhanced by congruent visual speech at a temporal scale that corresponds to the rate of syllables. There was also a significant effect at 16–18 Hz ($t_{(20)} = 3.8$, $p = 0.001$), although this finding is less compelling given the low reconstruction SNR at this frequency range.

A related question is whether we can ascertain which temporal scales are optimal for reconstructing the acoustic envelope from visual speech data. Addressing this issue is not entirely straightforward because there are many visual speech features at different levels of temporal granularity that correlate with the acoustic envelope (Jiang et al., 2002; Chandrasekaran et al., 2009; Jiang and Bernstein, 2011). In the stimulus reconstruction approach, the model reflects not only activity from auditory cortex that tracks the dynamics of the acoustic envelope, but also activity from potentially any visual area whose activity is correlated with the acoustic envelope and reflected in the EEG (Luo et al., 2010). Indeed, the reconstruction model can also indirectly index activity in brain areas whose activity is correlated with the acoustic

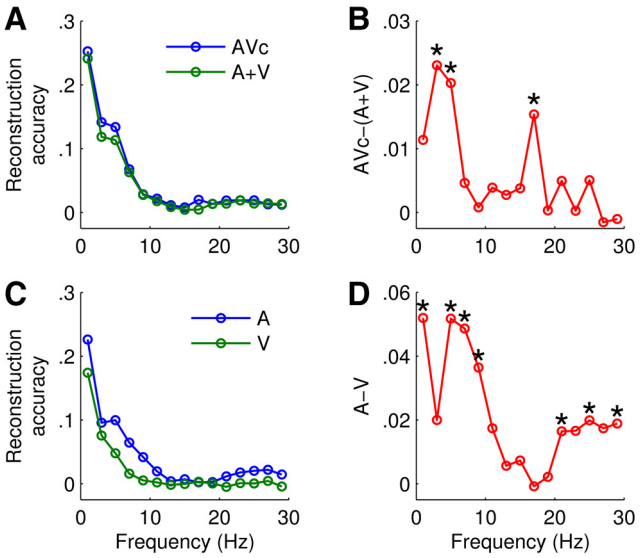

Figure 3. Reconstruction of the speech envelope from EEG at different temporal scales. ***A***, Mean ($N = 21$) reconstruction accuracy as a function of envelope frequency for the AVc (blue) and A+V (green) models. ***B***, Multisensory interaction effect [AVc − (A + V)] at each frequency band (*$p < 0.05$, *t* tests, Holm–Bonferroni corrected). ***C***, Mean ($N = 21$) reconstruction accuracy as a function of envelope frequency for the A (blue) and V (green) models. ***D***, Differences in unimodal model performance (A − V) at each frequency band (*$p < 0.05$, *t* tests, Holm–Bonferroni corrected).

envelope, even if that activity is not reflected directly in the data (Mesgarani et al., 2009). In one way, this is an advantage of the approach in that it is sensitive to visual speech processing without having to explicitly define specific visual speech features. However, it also makes it very difficult to tease apart the details of those visual speech contributions.

Bearing this in mind, we examined which frequencies optimized reconstruction of the acoustic envelope from the V data and compared it with those that optimized reconstruction using the A data (Fig. 3*C*). Reconstruction accuracy was significantly higher in the A condition at almost every frequency band ($p < 0.05$, *t* tests, Holm–Bonferroni corrected; Fig. 3*D*) except at two distinct spectral regions that, interestingly, corresponded to the two peaks in multisensory enhancement (2–4 Hz: $t_{(20)} = 1.8$, $p = 0.08$; 16–18 Hz: $t_{(20)} = 0.17$, $p = 0.87$).

Spatiotemporal profile of neuronal multisensory effects

To examine the temporal profile of our neuronal multisensory effects, we determined the temporal response function for each of the seven conditions, as well as the sum of the unimodal TRFs (A+V). Figure 4*A* shows the temporal profile of the TRFs for the congruent speech conditions at frontal channel Fz (top) and occipital channel Oz (bottom), whereas Figure 4*B* shows the TRFs for the incongruent speech conditions at the same channel locations. Comparing AVc with A+V as before, we see multisensory interaction effects in the form of a reduced amplitude over occipital scalp at ~140 ms (Oz: $t_{(20)} = 2.9$, $p = 0.01$; Fig. 4*C*, top) and over frontal scalp at ~220 ms (Fz: $t_{(20)} = 3.1$, $p = 0.006$; Fig. 4*C*, bottom).

To relate this late neuronal multisensory effect back to our stimulus reconstruction results, we examined the relative channel weightings of each of the reconstruction models. The channel weights represent the amount of information that each channel provides for reconstruction, i.e., highly informative channels receive weights of greater magnitude whereas channels providing little or no information receive weights closer to zero. However, unlike TRF model parameters, significant nonzero weights may also be observed at channels where cortical activity is statistically independent of stimulus tracking; hence, the spatiotemporal distribution of such model weights can be difficult to interpret in terms of underlying neural generators (Ding and Simon, 2012; Haufe et al., 2014).

Figure 4*D* shows the channel weighting for each model averaged over time lags that correspond to our neuronal multisensory effects (125–250 ms). Although not necessarily reflective of the underlying neural generators, the model weights clearly maintain distinct topographic patterns subject to stimulus modality. Channels over left and right temporal scalp make large contributions to stimulus reconstruction in the A model, whereas channels over occipital scalp are dominant in the V model. Unsurprisingly, channels over both temporal and occipital scalp make significant contributions in the congruent AV model, whereas only channels over temporal scalp make significant contributions in the incongruent AV models. This is because the incongruent visual stimuli were not informative of the acoustic envelope dynamics. The A+V model places significant weight on channels over temporal and occipital scalp, similar to the AVc model.

Although the AVc and A+V models appeared to have similar channel weightings, their ability to decode the speech envelope was significantly different. To better visualize the similarity relationships across all eight models, we represented channel weight dissimilarity in a two-dimensional Euclidean space using nonmetric MDS. Model dissimilarity was examined within two specific time intervals: an early interval (0–125 ms; Fig. 4*E*, left), at which latencies there were no multisensory effects evident in our TRF measures, and a later interval (125–250 ms; Fig. 4*E*, right), at which latencies there were significant multisensory effects evident in the TRFs. Visual inspection of the MDS plot for the earlier time interval (Fig. 4*E*, left) suggests that the models were organized into two discrete groupings consisting of audio and nonaudio stimuli. The AVc model is not visually discriminable from the other audio conditions at this interval, in line with the TRFs. In the later interval, however (Fig. 4*E*, right), the AVc model shows the greatest discriminability relative to the other models, indicating that it is capturing neuronal contributions from cross-modal interactions that are not well represented in the A+V model, also in agreement with the TRF results.

Discussion

We have demonstrated that when visual speech is congruent with auditory speech, the cortical representation of the speech envelope is enhanced relative to that predicted by the additive model criterion. These cross-modal interactions were most prominent at time scales indicative of syllabic integration (2–6 Hz). This was reflected in the neural responses by a suppression in amplitude at ~140 and ~220 ms, which corresponded with a late shift in the spatiotemporal profile of our reconstruction models, suggesting the involvement of neural generators that were not strongly activated during unimodal speech.

AV congruency and envelope tracking

Whereas envelope tracking was enhanced by congruent AV speech, it was inhibited when the A and V streams were incongruent both temporally and contextually (Fig. 2*B*). A possible explanation for this is that the male speaker's voice becomes less relevant when the visual stimulus is a female speaker or nature scene; hence, attentional resources dedicated to the auditory stimulus may have been reduced, a situation that is known to affect speech tracking (Ding and Simon,

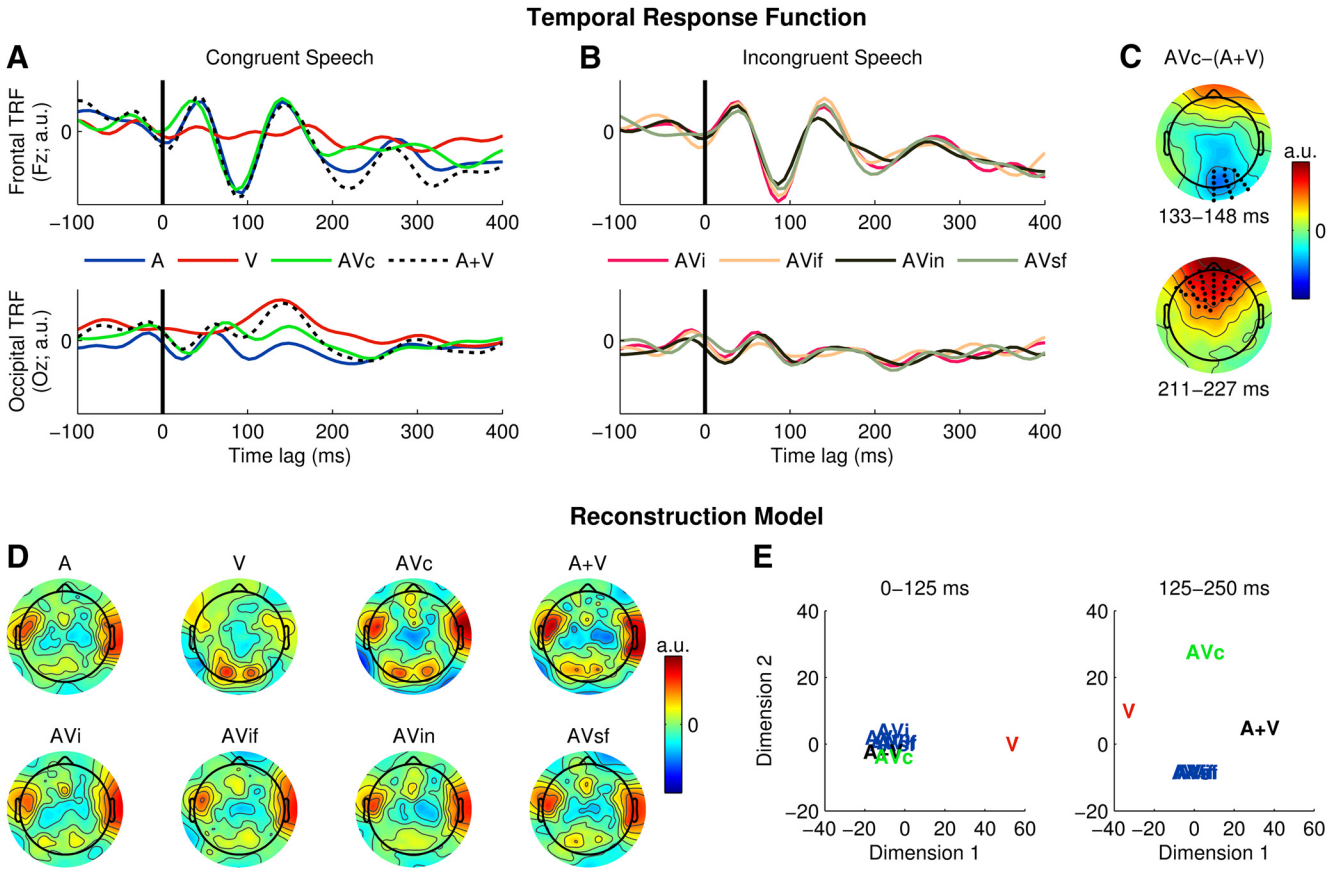

Figure 4. Spatiotemporal analysis of neuronal multisensory effects. ***A***, Group-average ($N = 21$) TRFs for congruent speech conditions at frontal scalp location Fz (top) and occipital scalp location Oz (bottom). ***B***, TRFs for incongruent speech conditions at the same scalp locations as in ***A***. ***C***, Topographic maps of multisensory interaction effects [AV − (A + V)] at ~140 ms (top) and ~220 ms (bottom). The black markers indicate channels where the interaction effect was significant across subjects ($p < 0.05$, t tests). ***D***, Group-average ($N = 21$) reconstruction models highlighting differential channel weightings at time lags corresponding to neuronal multisensory effects in ***C*** (125–250 ms). ***E***, Visualization of filter weight dissimilarity in a two-dimensional Euclidean space obtained using nonmetric multidimensional scaling for time lags between 0–125 ms (left) and 125–250 ms (right). Coloring was applied to highlight discrete groupings based on the 125–250 ms interval.

2012; O'Sullivan et al., 2015). This notion also fits with the theory that during conflicting AV presentation such as the McGurk scenario, directing attention toward a particular modality tends to reduce the bias of the unattended modality (Welch and Warren, 1980; Talsma et al., 2010).

We showed that the enhanced cortical entrainment in the case of the AVc condition exceeded that predicted by the additive model (Fig. 2*B*). This fits with recent views suggesting that visual speech increases the accuracy with which auditory cortex tracks the ongoing speech signal, leading to improved speech perception (Schroeder et al., 2008; Peelle and Sommers, 2015). However, our finding contrasts with a recent study (Zion Golumbic et al., 2013) that did not demonstrate enhanced neural tracking for single-speaker AV speech (but did for competing speakers). However, their finding was based on intertrial coherence, an indirect measure of envelope tracking, whereas stimulus reconstruction and TRF estimation are direct measures and, as such, may be more sensitive to subtle differences in tracking elicited during single-speaker AV speech. Furthermore, their stimuli were shorter (~10 s) and were repeated more times (40 per condition), meaning that the contribution of the visual stimulus may have varied based on the ability of participants to predict the upcoming auditory information.

Indeed, the effects of being able to predict the acoustic information may also be reflected in the reports by Zion Golumbic et al. (2013) of an early enhancement in TRF amplitude at ~50 ms (AV vs A). In contrast, we found that TRF amplitude was reduced at the later latencies of ~140 and ~220 ms (AVc vs A+V; Fig. 4*C*), in line with previous studies that have demonstrated emergent multisensory interactions in the form of suppressed cortical measures at ~120–190 ms (Besle et al., 2004b) and 160–220 ms (Bernstein et al., 2008). In keeping with recent perspectives on AV speech processing (Peelle and Sommers, 2015), we posit that this late suppression of cortical activity is reflective of an emergent integration stage that utilizes the relevant visual speech information to constrain the number of possible candidates. Indeed, this notion that emergent neuronal contributions may be driving our multisensory effects was also supported by our MDS analysis of the reconstruction models that revealed differential AVc versus A+V weight patterns only at later time lags (125–250 ms; Fig. 4*E*, right). It has been suggested (Peelle and Sommers, 2015) that earlier integration effects are likely reflective of increased auditory cortical sensitivity to acoustic information; thus, we predict that they may be more evident in complementary modes of AV speech such as speech-in-noise.

AV speech integration at the syllabic time scale

Our data suggest that envelope tracking is enhanced by congruent visual speech at a temporal scale that corresponds to the rate of syllables (2–6 Hz; Fig. 3*B*). This fits very well with a recent MEG study by Luo et al. (2010), which used natural, continuous AV stimuli to demonstrate that the phase of auditory cortex

tracks both auditory and visual stimulus dynamics and that this cross-modal phase modulation is most prominent in low-frequency neural information in the delta–theta band (2–7 Hz). This also fits with recent data that demonstrated a temporal correspondence between facial movements and the speech envelope in the 2–7 Hz frequency range (Chandrasekaran et al., 2009). Interestingly, there was no significant difference in the contribution from visual and auditory speech at frequencies where multisensory integration peaked (Fig. 3*D*). This may suggest that multisensory integration is enhanced for temporal scales where neither modality is particularly dominant, or at least where visual speech provides complementary information.

Future paradigms involving manipulations to the SNR of both the acoustic signal (e.g., speech-in-noise) and the visual signal (e.g., use of point light stimuli, dynamic annulus stimuli, and partially occluded faces) may lead to shifts in the spectral profile of the multisensory effects (Fig. 3*B*) and/or the unisensory effects (Fig. 3*D*), allowing firmer conclusions to be drawn. This endeavor might be aided further by extending the framework to reduce the reliance on the acoustic envelope by directly incorporating information about phonemes and visemes as has been done recently for unimodal auditory speech (Di Liberto et al., 2015). In addition, using other approaches to quantify AV correlations such as those based on mutual information models (Nock et al., 2002) and hidden Markov models (Rabiner, 1989) may provide important complementary insights.

Temporal coherence as a theoretical framework for AV integration

It has been suggested that the integration of auditory and visual speech could be driven by the temporal coherence of cross-modal information (Zion Golumbic et al., 2013). Computational and theoretical perspectives on stream segregation suggest that multifeature auditory sources are segregated into perceptual objects based on the temporal coherence of the neuronal responses to the various acoustic features (Elhilali et al., 2009; Shamma et al., 2011; Ding and Simon, 2012). Recently, Ding et al. (2014) demonstrated that cortical entrainment to the speech envelope does not reflect encoding of the envelope per se, as it relies on the spectrotemporal fine structure of speech. They suggest that it may instead index an analysis-by-synthesis mechanism, whereby spectrotemporal features that are correlated with the envelope are encoded during the synthesis phase (for review, see Ding and Simon, 2014). In keeping with previous work espousing a correlated mode of AV speech (Campbell, 2008), we postulate that visual speech features, being correlated with the envelope, results in the visual signal being bound to the auditory features to form a multisensory object.

Brain regions and neural mechanisms in AV integration

In terms of what specific neural populations might facilitate the binding of temporally coherent visual and auditory speech, one candidate region is the superior temporal sulcus, which has previously been linked with multisensory object formation (Calvert and Campbell, 2003; Beauchamp et al., 2004; Kayser and Logothetis, 2009). 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). That said, that our results may derive from emergent activity during AV speech could suggest a role for the supramarginal and angular gyrus (Bernstein et al., 2008), although that previous study found these effects only in the left hemisphere. Of course, in addition to such putatively multisensory regions, it remains a possibility that information pertaining to the timing of cross-modal stimuli could be projected to classic sensory-specific regions in a thalamocortical feedforward manner or laterally from other sensory-specific regions (Besle et al., 2008; Schroeder et al., 2008; Arnal et al., 2009). The latency of our multisensory effects may make this explanation less likely however, at least in the context of a correlated mode of AV speech.

A possible neural mechanism recently proposed also relates to the correlation between the speech envelope and visual motion. This theory 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 MEG work, which has demonstrated that the phase of oscillations in auditory cortex tracks the temporal structure of continuous visual speech (Luo et al., 2010), and fMRI work, which has demonstrated that the source of the visual facilitation of speech arises from motion-sensitive cortex (Arnal et al., 2009). Another suggestion for how visual speech may impact auditory speech processing is that this interaction may 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 incoming information at the phonetic level. This process could in part be mediated by preceding visual cues that could continually update auditory cortex before the arrival of such information.

An analysis-by-synthesis perspective of visual speech

We demonstrated that it was possible to reconstruct an estimate of the acoustic envelope from visual speech data with accuracy well above chance level (Fig. 2*B*). Although the acoustic envelope was not explicitly encoded in the neural data during visual speech, it may still be inferred if some correlated feature of the visual speech was encoded (Mesgarani et al., 2009), as discussed above. One possible explanation is that instantaneous measures of motion during visual speech are highly correlated with the amplitude of the acoustic envelope (Chandrasekaran et al., 2009). However, in keeping with an analysis-by-synthesis framework, Crosse et al. (2015) suggest that such occipital activity may, in fact, reflect the processing of higher-level visual speech features in visual cortex in addition to just motion tracking. It has been demonstrated that every level of speech structure can be perceived visually, thus suggesting that there are visual modality-specific representations of speech in visual brain areas and not just in auditory brain areas (for review, see Bernstein and Liebenthal, 2014). Furthermore, we observed a strong correlation between behavior and envelope tracking in the visual speech data (Fig. 2*C*), similar to that recently demonstrated in auditory speech-in-noise (Ding and Simon, 2013). As such, we tentatively posit that lipreading accuracy is reflected in the neural tracking of the envelope and that this tracking process includes the synthesis of visual speech tokens in visual-specific brain regions. Whereas we have outlined above the challenges associated with using stimulus reconstruction to tease this issue apart, the use of different paradigms, such as those mentioned above, within our framework may yet prove enlightening.

Summary and conclusions

We have established a framework for investigating multisensory integration in the context of natural, continuous speech. This naturalistic approach 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., 2013), and schizophrenia (Ross et al., 2007b; Stekelenburg et al., 2013). Although it will certainly require methods complementary to EEG to determine the details of the neural mechanisms underlying AV speech integration, we suggest that the effects reported here are mediated by the temporal coherence of congruent AV speech at the syllabic time scale as part of an analysis-by-synthesis process (Ding and Simon, 2014). Future work examining this using a speech-in-noise paradigm should prove more informative given the well-established benefits of multisensory speech in adverse hearing conditions (Sumby and Pollack, 1954; Ross et al., 2007a).

References

Abrams DA, Nicol T, Zecker S, Kraus N (2008) Right-hemisphere auditory cortex is dominant for coding syllable patterns in speech. J Neurosci 28:3958–3965. CrossRef Medline

Ahissar E, Nagarajan S, Ahissar M, Protopapas A, Mahncke H, Merzenich MM (2001) Speech comprehension is correlated with temporal response patterns recorded from auditory cortex. Proc Natl Acad Sci U S A 98:13367–13372. CrossRef Medline

Arnal LH, Morillon B, Kell CA, Giraud AL (2009) Dual neural routing of visual facilitation in speech processing. J Neurosci 29:13445–13453. CrossRef Medline

Beauchamp MS, Lee KE, Argall BD, Martin A (2004) Integration of auditory and visual information about objects in superior temporal sulcus. Neuron 41:809–823. CrossRef Medline

Bernstein LE, Liebenthal E (2014) Neural pathways for visual speech perception. Front Neurosci 8:386. CrossRef Medline

Bernstein LE, Auer ET Jr, Wagner M, Ponton CW (2008) Spatiotemporal dynamics of audiovisual speech processing. Neuroimage 39:423–435. CrossRef Medline

Besle J, Fort A, Giard M-H (2004a) Interest and validity of the additive model in electrophysiological studies of multisensory interactions. Cogn Process 5:189–192.

Besle J, Fort A, Delpuech C, Giard MH (2004b) Bimodal speech: early suppressive visual effects in human auditory cortex. Eur J Neurosci 20:2225–2234. CrossRef Medline

Besle J, Fischer C, Bidet-Caulet A, Lecaignard F, Bertrand O, Giard MH (2008) Visual activation and audiovisual interactions in the auditory cortex during speech perception: intracranial recordings in humans. J Neurosci 28:14301–14310. CrossRef Medline

Brandwein AB, Foxe JJ, Butler JS, Russo NN, Altschuler TS, Gomes H, Molholm S (2013) The development of multisensory integration in high-functioning autism: high-density electrical mapping and psychophysical measures reveal impairments in the processing of audiovisual inputs. Cereb Cortex 23:1329–1341. CrossRef Medline

Calvert GA, Campbell R (2003) Reading speech from still and moving faces: the neural substrates of visible speech. J Cogn Neurosci 15:57–70. CrossRef Medline

Campbell R (2008) The processing of audio-visual speech: empirical and neural bases. Philos Trans R Soc Lond B Biol Sci 363:1001–1010. CrossRef Medline

Chandrasekaran C, Trubanova A, Stillittano S, Caplier A, Ghazanfar AA (2009) The natural statistics of audiovisual speech. PLoS Comput Biol 5:e1000436. CrossRef Medline

Chang EF, Rieger JW, Johnson K, Berger MS, Barbaro NM, Knight RT (2010) Categorical speech representation in human superior temporal gyrus. Nat Neurosci 13:1428–1432. CrossRef Medline

Crosse MJ, ElShafei HA, Foxe JJ, Lalor EC (2015) Investigating the temporal dynamics of auditory cortical activation to silent lipreading. Paper presented at 7th International IEEE/EMBS Conference on Neural Engineering, Montpellier, France, April.

Delorme A, Makeig S (2004) EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. J Neurosci Methods 134:9–21. CrossRef Medline

Di Liberto GM, O'Sullivan JA, Lalor EC (2015) Low frequency cortical entrainment to speech reflects phonemic level processing. Curr Biol. Advance online publication. Retrieved Sept. 24, 2015. CrossRef

Ding N, Simon JZ (2012) Emergence of neural encoding of auditory objects while listening to competing speakers. Proc Natl Acad Sci U S A 109:11854–11859. CrossRef Medline

Ding N, Simon JZ (2013) Adaptive temporal encoding leads to a background-insensitive cortical representation of speech. J Neurosci 33:5728–5735. CrossRef Medline

Ding N, Simon JZ (2014) Cortical entrainment to continuous speech: functional roles and interpretations. Front Hum Neurosci 8:311. CrossRef Medline

Ding N, Chatterjee M, Simon JZ (2014) Robust cortical entrainment to the speech envelope relies on the spectro-temporal fine structure. Neuroimage 88:41–46. CrossRef Medline

Elhilali M, Ma L, Micheyl C, Oxenham AJ, Shamma SA (2009) Temporal coherence in the perceptual organization and cortical representation of auditory scenes. Neuron 61:317–329. CrossRef Medline

Giraud AL, Poeppel D (2012) Cortical oscillations and speech processing: emerging computational principles and operations. Nat Neurosci 15:511–517. CrossRef Medline

Grant KW, Seitz PF (2000) The use of visible speech cues for improving auditory detection of spoken sentences. J Acoust Soc Am 108:1197–1208. CrossRef Medline

Grant KW, Walden BE, Seitz PF (1998) Auditory-visual speech recognition by hearing-impaired subjects: consonant recognition, sentence recognition, and auditory-visual integration. J Acoust Soc Am 103:2677–2690. CrossRef Medline

Hairston WD, Burdette JH, Flowers DL, Wood FB, Wallace MT (2005) Altered temporal profile of visual-auditory multisensory interactions in dyslexia. Exp Brain Res 166:474–480. CrossRef Medline

Haufe S, Meinecke F, Görgen K, Dähne S, Haynes JD, Blankertz B, Bießmann F (2014) On the interpretation of weight vectors of linear models in multivariate neuroimaging. Neuroimage 87:96–110. CrossRef Medline

Jiang J, Bernstein LE (2011) Psychophysics of the McGurk and other audiovisual speech integration effects. J Exp Psychol Hum Percept Perform 37:1193–1209. CrossRef Medline

Jiang J, Alwan A, Keating PA, Auer ET, Bernstein LE (2002) On the relationship between face movements, tongue movements, and speech acoustics. EURASIP J Appl Signal Process 11:1174–1188.

Kayser C, Logothetis NK (2009) Directed interactions between auditory and superior temporal cortices and their role in sensory integration. Front Integr Neurosci 3:7. CrossRef Medline

Kayser C, Petkov CI, Logothetis NK (2008) Visual modulation of neurons in auditory cortex. Cereb Cortex 18:1560–1574. CrossRef Medline

Klucharev V, Möttönen R, Sams M (2003) Electrophysiological indicators of phonetic and non-phonetic multisensory interactions during audiovisual speech perception. Cogn Brain Res 18:65–75. CrossRef Medline

Kruskal JB, Wish M (1978) Multidimensional scaling. Newbury Park, CA: Sage.

Lalor EC, Foxe JJ (2010) Neural responses to uninterrupted natural speech can be extracted with precise temporal resolution. Eur J Neurosci 31:189–193. CrossRef Medline

Luo H, Liu Z, Poeppel D (2010) Auditory cortex tracks both auditory and visual stimulus dynamics using low-frequency neuronal phase modulation. PLoS Biol 8:e1000445. CrossRef Medline

McGurk H, MacDonald J (1976) Hearing lips and seeing voices. Nature 264:746–748. CrossRef Medline

Mesgarani N, David SV, Fritz JB, Shamma SA (2009) Influence of context and behavior on stimulus reconstruction from neural activity in primary auditory cortex. J Neurophysiol 102:3329–3339. CrossRef Medline

Miller J (1982) Divided attention: evidence for coactivation with redundant signals. Cogn Psychol 14:247–279. CrossRef Medline

Möttönen R, Krause CM, Tiippana K, Sams M (2002) Processing of changes in visual speech in the human auditory cortex. Cogn Brain Res 13:417–425. CrossRef Medline

Nock HJ, Iyengar G, Neti C (2002) Assessing face and speech consistency for monologue detection in video. Paper presented at 10th ACM International Conference on Multimedia, Juan-les-Pins, France, December.

O'Sullivan JA, Crosse MJ, Power AJ, Lalor EC (2013) The effects of attention

and visual input on the representation of natural speech in EEG. Paper presented at 35th Annual International Conference of the IEEE/EMBS, Osaka, Japan, July.

O'Sullivan JA, Power AJ, Mesgarani N, Rajaram S, Foxe JJ, Shinn-Cunningham BG, Slaney M, Shamma SA, Lalor EC (2015) Attentional selection in a cocktail party environment can be decoded from single-trial EEG. Cereb Cortex 25:1697–1706. CrossRef Medline

Pasley BN, David SV, Mesgarani N, Flinker A, Shamma SA, Crone NE, Knight RT, Chang EF (2012) Reconstructing speech from human auditory cortex. PLoS Biol 10:e1001251. CrossRef Medline

Peelle JE, Sommers MS (2015) Prediction and constraint in audiovisual speech perception. Cortex 68:169–181. CrossRef Medline

Power AJ, Foxe JJ, Forde EJ, Reilly RB, Lalor EC (2012) At what time is the cocktail party? A late locus of selective attention to natural speech. Eur J Neurosci 35:1497–1503. CrossRef Medline

Raab DH (1962) Statistical facilitation of simple reaction times. Trans N Y Acad Sci 24:574–590. CrossRef Medline

Rabiner LR (1989) A tutorial on hidden Markov models and selected applications in speech recognition. Proc IEEE 77:257–286. CrossRef

Reisberg D, McLean J, Goldfield A (1987) Easy to hear but hard to understand: a lip-reading advantage with intact auditory stimuli. In: Hearing by eye: the psychology of lip-reading (Dodd B, Campbell R, eds), pp 97–114. Hillsdale, NJ: Erlbaum.

Rieke F, Bodnar D, Bialek W (1995) Naturalistic stimuli increase the rate and efficiency of information transmission by primary auditory afferents. Philos Trans R Soc Lond B Biol Sci 262:259–265. CrossRef Medline

Rosen S (1992) Temporal information in speech: acoustic, auditory and linguistic aspects. Philos Trans R Soc Lond B Biol Sci 336:367–373. CrossRef Medline

Ross LA, Saint-Amour D, Leavitt VM, Javitt DC, Foxe JJ (2007a) Do you see what I am saying? Exploring visual enhancement of speech comprehension in noisy environment. Cereb Cortex 17:1147–1153. CrossRef Medline

Ross LA, Saint-Amour D, Leavitt VM, Molholm S, Javitt DC, Foxe JJ (2007b) Impaired multisensory processing in schizophrenia: deficits in the visual enhancement of speech comprehension under noisy environmental conditions. Schizophr Res 97:173–183. CrossRef Medline

Sams M, Aulanko R, Hämäläinen M, Hari R, Lounasmaa OV, Lu ST, Simola J (1991) Seeing speech: visual information from lip movements modifies activity in the human auditory cortex. Neurosci Lett 127:141–145. CrossRef Medline

Schroeder CE, Lakatos P, Kajikawa Y, Partan S, Puce A (2008) Neuronal oscillations and visual amplification of speech. Trends Cogn Sci 12:106–113. CrossRef Medline

Shamma SA, Elhilali M, Micheyl C (2011) Temporal coherence and attention in auditory scene analysis. Trends Neurosci 34:114–123. CrossRef Medline

Sheedy CM, Power AJ, Reilly RB, Crosse MJ, Loughnane GM, Lalor EC (2014) Endogenous auditory frequency-based attention modulates electroencephalogram-based measures of obligatory sensory activity in humans. Neuroreport 25:219–225. CrossRef Medline

Stein BE, Meredith MA (1993) The merging of the senses. Cambridge, MA: MIT.

Stekelenburg JJ, Vroomen J (2007) Neural correlates of multisensory integration of ecologically valid audiovisual events. J Cogn Neurosci 19:1964–1973. CrossRef Medline

Stekelenburg JJ, Maes JP, Van Gool AR, Sitskoorn M, Vroomen J (2013) Deficient multisensory integration in schizophrenia: an event-related potential study. Schizophr Res 147:253–261. CrossRef Medline

Sumby WH, Pollack I (1954) Visual contribution to speech intelligibility in noise. J Acoust Soc Am 26:212–215. CrossRef

Summerfield Q (1987) Some preliminaries to a comprehensive account of audio-visual speech perception. Hillsdale, NJ: Erlbaum.

Summerfield Q (1992) Lipreading and audio-visual speech perception. Philos Trans R Soc Lond B Biol Sci 335:71–78. CrossRef Medline

Talsma D, Senkowski D, Soto-Faraco S, Woldorff MG (2010) The multifaceted interplay between attention and multisensory integration. Trends Cogn Sci 14:400–410. CrossRef Medline

Teki S, Chait M, Kumar S, von Kriegstein K, Griffiths TD (2011) Brain bases for auditory stimulus-driven figure-ground segregation. J Neurosci 31:164–171. CrossRef Medline

Tjan BS, Chao E, Bernstein LE (2014) A visual or tactile signal makes auditory speech detection more efficient by reducing uncertainty. Eur J Neurosci 39:1323–1331. CrossRef Medline

Ulrich R, Miller J, Schröter H (2007) Testing the race model inequality: an algorithm and computer programs. Behav Res Methods 39:291–302. CrossRef Medline

van Wassenhove V, Grant KW, Poeppel D (2005) Visual speech speeds up the neural processing of auditory speech. Proc Natl Acad Sci U S A 102:1181–1186. CrossRef Medline

Welch RB, Warren DH (1980) Immediate perceptual response to intersensory discrepancy. Psychol Bull 88:638–667. CrossRef Medline

Zion Golumbic EM, Cogan GB, Schroeder CE, Poeppel D (2013) Visual input enhances selective speech envelope tracking in auditory cortex at a "cocktail party." J Neurosci 33:1417–1426. CrossRef Medline