Systems/Circuits

Eye Can Hear Clearly Now: Inverse Effectiveness in Natural Audiovisual Speech Processing Relies on Long-Term Crossmodal Temporal Integration

Michael J. Crosse, Giovanni M. Di Liberto, and Edmund C. Lalor
School of Engineering, Trinity Centre for Bioengineering, and Trinity College Institute of Neuroscience, Trinity College Dublin, Dublin 2, Ireland

Speech comprehension is improved by viewing a speaker's face, especially in adverse hearing conditions, a principle known as inverse effectiveness. However, the neural mechanisms that help to optimize how we integrate auditory and visual speech in such suboptimal conversational environments are not yet fully understood. Using human EEG recordings, we examined how visual speech enhances the cortical representation of auditory speech at a signal-to-noise ratio that maximized the perceptual benefit conferred by multisensory processing relative to unisensory processing. We found that the influence of visual input on the neural tracking of the audio speech signal was significantly greater in noisy than in quiet listening conditions, consistent with the principle of inverse effectiveness. Although envelope tracking during audio-only speech was greatly reduced by background noise at an early processing stage, it was markedly restored by the addition of visual speech input. In background noise, multisensory integration occurred at much lower frequencies and was shown to predict the multisensory gain in behavioral performance at a time lag of ~250 ms. Critically, we demonstrated that inverse effectiveness, in the context of natural audiovisual (AV) speech processing, relies on crossmodal integration over long temporal windows. Our findings suggest that disparate integration mechanisms contribute to the efficient processing of AV speech in background noise.

Key words: **EEG; envelope tracking; multisensory integration; speech intelligibility; speech-in-noise; stimulus reconstruction**

Significance Statement

The behavioral benefit of seeing a speaker's face during conversation is especially pronounced in challenging listening environments. However, the neural mechanisms underlying this phenomenon, known as inverse effectiveness, have not yet been established. Here, we examine this in the human brain using natural speech-in-noise stimuli that were designed specifically to maximize the behavioral benefit of audiovisual (AV) speech. We find that this benefit arises from our ability to integrate multimodal information over longer periods of time. Our data also suggest that the addition of visual speech restores early tracking of the acoustic speech signal during excessive background noise. These findings support and extend current mechanistic perspectives on AV speech perception.

Introduction

It has long been established that the behavioral benefits of audiovisual (AV) speech are more apparent in acoustic conditions in which intelligibility is reduced (Sumby and Pollack, 1954; Erber, 1975; Grant and Seitz, 2000; Bernstein et al., 2004; Ross et al., 2007). Enhanced multisensory processing in response to weaker sensory inputs is a phenomenon known as inverse effectiveness (Meredith and Stein, 1986). However, in the context of AV speech processing, there are particular audio signal-to-noise ratios (SNRs) at which the benefits of multisensory processing become maximized—a sort of multisensory "sweet spot" (Ross et al., 2007; Ma et al., 2009). It is likely that, when processing AV speech in such conditions, the brain must exploit both correlated and complementary visual information to optimize intelligibility (Summerfield, 1987; Grant and Seitz, 2000; Campbell, 2008). This could be achieved through multiple integration mechanisms occurring at different temporal stages. Specifically, recent perspectives on multistage AV speech processing suggest that visual speech provides cues to the timing of the acoustic signal that could project directly from visual cortex, increasing the sensitiv-

Received April 25, 2016; revised July 12, 2016; accepted Aug. 3, 2016.

Author contributions: M.J.C. and E.C.L. designed research; M.J.C. performed research; M.J.C., G.M.D.L., and E.C.L. contributed unpublished reagents/analytic tools; M.J.C. and G.M.D.L. analyzed data; M.J.C. and E.C.L. wrote the paper.

This work was supported by the Programme for Research in Third-Level Institutions and cofounded under the European Regional Development fund.

The authors declare no competing financial interests.

Correspondence should be addressed to Edmund C. Lalor, Ph.D., Department of Biomedical Engineering, 201 Robert B. Goergen Hall, P.O. Box 270168, Rochester, NY 14627. E-mail: edmund_lalor@urmc.rochester.edu.

M.J. Crosse's present affiliation: Department of Pediatrics and Department of Neuroscience, Albert Einstein College of Medicine, Bronx, New York 10461.

DOI:10.1523/JNEUROSCI.1396-16.2016

 0270-6474/16/369888-08$15.00/0

ity of auditory cortex to the upcoming acoustic information, whereas complementary visual cues that convey place and manner of articulation could be integrated with converging acoustic information in supramodal regions such as superior temporal sulcus (STS), serving to constrain lexical selection (Peelle and Sommers, 2015).

Studying how the brain uses the timing and lexical constraints of visual speech to enhance the processing of acoustic information necessitates the use of natural, conversation-like speech stimuli. Recent EEG and MEG studies have used naturalistic speech stimuli to examine how visual speech effects the cortical representation of the speech envelope (Zion Golumbic et al., 2013; Crosse et al., 2015). However, it is not yet known how these neural measures of speech processing are affected by visual speech at much lower SNRs, where multisensory processing is optimized. In particular, the specific neural mechanisms invoked in such situations are poorly understood. A recent MEG study examined how different levels of noise affect the cortical representation of audio-only speech and demonstrated that it is relatively insensitive to background noise, even at low SNRs at which intelligibility is diminished (Ding and Simon, 2013). Only when intelligibility reached the perithreshold level (e.g., at an SNR of −9 dB) did they find that envelope tracking was significantly reduced. Given that AV speech has been shown to improve intelligibility in noise equivalent to an increase in SNR of up to 15 dB (Sumby and Pollack, 1954), we hypothesized that the addition of visual cues could substantially restore envelope tracking in such perithreshold conditions.

Here, an AV speech-in-noise paradigm was implemented to study the neural interaction between continuous auditory and visual speech at an SNR at which multisensory processing was of maximal benefit relative to unisensory processing. High-density EEG recordings were analyzed using a recently introduced system identification framework for indexing multisensory integration in natural AV speech (Crosse et al., 2015). We provide evidence that neural entrainment to continuous AV speech conforms to the principle of inverse effectiveness and that it does so specifically by restoring early tracking of the acoustic speech signal and integrating low-frequency crossmodal information over longer temporal windows. These findings support the notion that fundamentally different integration mechanisms contribute to the efficient processing of AV speech in adverse listening environments (Schwartz et al., 2004; van Wassenhove et al., 2005; Eskelund et al., 2011; Baart et al., 2014; Peelle and Sommers, 2015). Our results also suggest that in degraded listening environments, crossmodal integration of AV speech occurs at a more coarse-grained linguistic level.

Materials and Methods

To determine how AV speech processing is affected by SNR, we analyzed data from two separate experiments: a "speech-in-quiet" paradigm and a "speech-in-noise" paradigm, each of which used the same target detection task but involved separate participant samples.

Participants. Twenty-one participants (8 females; age range: 19–37 years) completed the speech-in-quiet experiment as part of a separate study (Crosse et al., 2015) and 21 different participants (6 females; age range: 21–35 years) completed the speech-in-noise experiment. All participants were native English speakers, had self-reported normal hearing and normal or corrected-to-normal vision, were free of neurological diseases, and provided written informed consent. All procedures were undertaken in accordance with the Declaration of Helsinki and were approved by the Ethics Committee of the Health Sciences Faculty at Trinity College Dublin.

Stimuli and procedure. The stimuli used in both experiments were drawn from a set of videos that consisted of an American English male speaking in a conversational-like manner. Fifteen 60-s videos were rendered into 1280 × 720-pixel movies at 30 frames/s and exported in audio-only (A), visual-only (V), and AV format in VideoPad Video Editor (NCH Software). The soundtracks were sampled at 48 kHz, underwent dynamic range compression, and were matched in intensity (as measured by root mean square; see Crosse et al., 2015). For the speech-in-noise experiment, the soundtracks were additionally mixed with spectrally matched stationary noise to ensure consistent masking across stimuli (Ding and Simon, 2013; Ding et al., 2014). The noise stimuli were generated in MATLAB (The MathWorks) using a 50th-order forward linear predictive model estimated from the original speech recording. Prediction order was calculated based on the sampling rate of the soundtracks (Parsons, 1987).

Behavioral piloting was used to select the SNR value such that it maximized the increase in intelligibility produced by AV speech relative to A speech. A subset of participants ($n = 3$) listened to four 60-s passages of A and AV speech at SNRs of −7, −9, and −11 dB. After each passage, they were asked to rate as a percentage how intelligible the speech was. These data indicated that an SNR of −9 dB yielded the largest perceptual gain and thus was chosen for the main experiment. The same spectrally matched noise stimuli were also presented in the V condition, but without any speech content.

In both experiments, EEG recording took place in a dark sound-attenuated room with participants seated 70 cm from the visual display. Stimulus presentation was controlled using Presentation software (Neurobehavioral Systems). Visual stimuli were presented at a refresh rate of 60 Hz on a 19-inch CRT monitor and audio stimuli were presented diotically through Sennheiser HD650 headphones at 48 kHz. The same target word detection task was used to encourage active engagement with the speech content in both experiments (Crosse et al., 2015). In addition to detecting target words, participants in the speech-in-noise experiment were required to rate subjectively the intelligibility of the speech stimuli at the end of each 60-s trial. Intelligibility was rated as a percentage of the total words understood using a 10-point scale (0–10%, 10–20%, ... 90–100%). Stimulus presentation order was completely random in the speech-in-quiet experiment; however, this approach was not suitable for the speech-in-noise paradigm because, if the same speech passage was presented twice in quick succession (albeit in different conditions), it could potentially influence intelligibility in the latter condition. Instead, the 15 passages were ordered 1–15 and presented 3 times, but the condition from trial-to-trial was randomized. In this way, each speech passage could not be repeated in another format within 15 trials of the preceding one.

Behavioral data analysis. To identify a behavioral measure of multisensory integration (MSI), we investigated whether the probability of detecting a multisensory stimulus exceeded the statistical facilitation produced by the unisensory stimuli. False-positives were accounted for by taking an F-measure of each participant's detection rate. F-scores (or F_1 scores) were calculated as the harmonic mean of precision and recall (Van Rijsbergen, 1979). Therefore, our behavioral MSI measure was calculated as follows:

$$\mathrm{MSI}_{\mathrm{Behav}} = F_1(\mathrm{AV}) - \hat{F}_1(\mathrm{AV}) \quad (1)$$

where $F_1(\mathrm{AV})$ is the F_1 score for the AV condition and $\hat{F}_1(\mathrm{AV})$ is the predicted F_1 score based on the values of the unisensory conditions. Although the same detection task was implemented in both experiments, two different criteria were used to quantify $\hat{F}_1(\mathrm{AV})$ as outlined in Stevenson et al. (2014). For speech-in-quiet, detection accuracy was near ceiling so a maximum criterion model was used: $\hat{F}_1(\mathrm{AV}) = \max[F_1(\mathrm{A}), F_1(\mathrm{V})]$. For speech-in-noise, accuracy was not at ceiling so a more conservative model was used that accounted for statistical facilitation (Blamey et al., 1989): $\hat{F}_1(\mathrm{AV}) = F_1(\mathrm{A}) + F_1(\mathrm{V}) - F_1(\mathrm{A}) \times F_1(\mathrm{V})$. Essentially, the term on the right represents the detection rate that would be expected when auditory and visual stimuli were presented together and processed independently (Stevenson et al., 2014). To quantify the gain in performance produced by AV speech, we calculated $\mathrm{MSI}_{\mathrm{Behav}}$ as a percentage of $\hat{F}_1(\mathrm{AV})$, in other words, as a percentage of independent unisensory processing.

EEG acquisition and preprocessing. In both experiments, 128-channel EEG data (plus mastoid channels) were acquired at a rate of 512 Hz using an ActiveTwo system (BioSemi). Triggers indicating the start of each trial were sent to the BioSemi system using an Arduino Uno microcontroller that detected an audio click at the start of each soundtrack by sampling the headphone output from the PC. Offline, the data were band-pass filtered between 0.3 and 30 Hz, downsampled to 64 Hz, and rereferenced to the average of the mastoid channels in MATLAB. To identify channels with excessive noise, the time series were visually inspected and the SD of each channel was compared with that of the surrounding channels. Channels contaminated by noise were recalculated by spline interpolating the surrounding clean channels in EEGLAB (Delorme and Makeig, 2004). Trials contaminated by excessive low-frequency noise were detrended using a sinusoidal function in NoiseTools (http://audition.ens.fr/adc/NoiseTools/).

Stimulus characterization. In this study, EEG analysis focused on the speech signal below 3 kHz because the strongest correlation between the mouth opening and vocal acoustics is between 2 and 3 kHz (Chandrasekaran et al., 2009; Grant and Seitz, 2000; Grant, 2001), meaning that visual speech can provide cues to the timing of less salient auditory events within this frequency range. Furthermore, visual speech can offer complementary information in the form of place of articulation, which can help to distinguish ambiguous acoustic content, particularly in second formant space.

The spectrogram representation of each stimulus was generated using a compressive gammachirp auditory filter bank that modeled the auditory periphery (Irino and Patterson, 2006). Outer and middle ear correction were applied using an FIR minimum phase filter before the stimuli were band-pass filtered into 256 logarithmically spaced frequency bands between 80 and 3000 Hz. The energy in each frequency band was calculated using a Hilbert transform and the broadband envelope was obtained by averaging across the frequency bands of the resulting spectrogram.

The rates of different linguistic units (e.g., words, syllables, vowels, consonants) in the speech stimuli were extracted from the audio files using the Forced Alignment and Vowel Extraction (FAVE) software suite (http://fave.ling.upenn.edu). This returns the start and end time points for individual phonemes, enabling detailed characterization of the timescale of both segmental and suprasegmental speech units.

Stimulus reconstruction. Neural tracking of the speech signal was measured in terms of how accurately the broadband speech envelope, $s(t)$, could be reconstructed from the EEG data, $r(t)$, using the following linear model:

$$\hat{s}(t) = \sum_{n=1}^{128} \sum_{\tau=0}^{500\text{ms}} r(t+\tau, n)\, g(\tau, n) \quad (2)$$

where $\hat{s}(t)$ is the estimated speech envelope, $r(t+\tau,n)$ is the EEG response at channel n and time lag τ, and $g(\tau,n)$ is the linear decoder for the corresponding channel and time lag. The objective was to reconstruct the underlying speech envelope (as opposed to the actual speech-in-noise mixture) because we only care about how the brain processes speech information. In any case, previous work has demonstrated that the underlying speech signal can be reconstructed from cortical activity with greater accuracy than the actual speech-in-noise mixture (Ding and Simon, 2013). The decoder $g(\tau,n)$ was optimized for each condition using ridge regression with leave-one-out cross-validation (Crosse et al., 2015; mTRF Toolbox; http://sourceforge.net/projects/aespa/) to maximize the correlation between $\hat{s}(t)$ and $s(t)$. As with the behavioral data, we define a neural measure of multisensory integration as follows:

$$\text{MSI}_{\text{EEG}} = \text{corr}[\hat{s}_{\text{AV}}(t), s(t)] - \text{corr}[\hat{s}_{\text{A+V}}(t), s(t)] \quad (3)$$

where $\hat{s}_{\text{AV}}(t)$ is the reconstructed envelope for the AV condition and $\hat{s}_{\text{A+V}}(t)$ is the estimated envelope for the additive unisensory model. Similar to the behavioral analysis, we defined multisensory gain by calculating MSI_{EEG} as a percentage of $\text{corr}[\hat{s}_{\text{A+V}}(t), s(t)]$.

Single-lag analysis. When reconstructing the speech envelope, the decoder $g(\tau,n)$ integrates EEG over a 500 ms window. This ensures that we capture important temporal information in the EEG that relates to each sample of the stimulus that we are trying to reconstruct. To quantify the contribution of each time lag toward reconstruction, decoders were trained on EEG at individual lags from 0 to 500 ms, instead of integrating across them (O'Sullivan et al., 2015). For a sampling frequency of 64 Hz, this equates to 33 individual lags and thus 33 separate decoders. For each time lag, the solution then becomes:

$$\hat{s}_{\tau}(t) = \sum_{n=1}^{128} r(t+\tau, n)\, g(\tau, n), \quad 0 < \tau \leq 500 \text{ ms} \quad (4)$$

where $\hat{s}_{\tau}(t)$ is the estimated speech envelope for lag τ. Because the decoders consisted of only a single time lag, there was no need for regularization along the time dimension. Instead of using ridge regression to compute the decoder, it was approximated by performing a singular value decomposition of the auto-correlation matrix (Theunissen et al., 2000; Mesgarani et al., 2009; Ding and Simon, 2012). Here, only those eigenvalues that exceed a specific fraction of the largest eigenvalue or peak power are included in the analysis. Qualitatively, this approach yields the same result as doing ridge regression. To examine how MSI_{EEG} varied as a function of time lag, it was calculated as before (Eq. 2) using the single-lag decoders. To investigate whether MSI_{EEG} was predictive of $\text{MSI}_{\text{Behav}}$ at a particular time lag, we calculated the correlation coefficient between the two measures across subjects. This was examined in speech-in-noise, where behavioral performance was not at ceiling.

Statistical analyses. All statistical analyses were conducted using two-way mixed ANOVAs with a between-subjects factor of SNR (quiet vs −9 dB) and a within-subjects factor of condition (A, V, A+V, AV), except where otherwise stated. Where sphericity was violated in factors with two or more levels, the Greenhouse–Geisser corrected degrees of freedom are reported. *Post hoc* comparisons were conducted using two-tailed t tests and multiple comparisons were corrected for using the Holm–Bonferroni method. All numerical values are reported as mean ± SD. Outlying participants were excluded from specific analyses if their values within that analysis were a distance of more than three times the interquartile range.

Results

Behavior and multisensory gain

Subjectively-rated intelligibility in the speech-in-noise experiment confirmed that intelligibility was highest in the AV condition ($t_{(20)} = 10.3$, $p = 1.9 \times 10^{-9}$; A+V: 36.9 ± 18.4%; AV: 63.6 ± 15.8%, Fig. 1*B*). This was reflected in how accurately participants could detect the target words, with detection accuracy significantly higher in the AV condition compared with that predicted by the unisensory scores ($t_{(20)} = 2.6$, $p = 0.018$; $\hat{F}_1$(AV): 0.7 ± 0.09, F_1(AV): 0.76 ± 0.08; Fig. 1*C*, left). In speech-in-quiet, accuracy in the A and AV conditions was at ceiling, so there was no observable multisensory benefit. As a result, the AV gain for speech-in-noise was significantly greater than that for speech-in-quiet [unpaired t test: $t_{(39)} = 2.8$, $p = 0.0086$; $\text{MSI}_{\text{Behav}}$ (quiet): −1.44 ± 5.61%, $\text{MSI}_{\text{Behav}}$ (−9 dB): 9.14 ± 15.12%; Fig. 1*C*, right]. For speech-in-noise, both intelligibility and detection accuracy varied substantially across subjects. Importantly, the individual accuracy scores were shown to be significantly correlated with intelligibility in both the unisensory conditions (A: $r = 0.51$, $p = 0.02$; V: $r = 0.55$, $p = 0.01$). In the AV condition, accuracy rates were nearer to ceiling, so any observable correlation with intelligibility was most likely obscured.

Neural enhancement and inverse effectiveness

Neural tracking of the speech signal was measured based on how accurately the broadband envelope could be reconstructed from the participants' EEG (Fig. 2*A*, left). A mixed ANOVA with factors of SNR (quiet vs −9 dB) and condition (A vs V) revealed a significant interaction effect ($F_{(1,40)} = 24.1$, $p = 1.6 \times 10^{-5}$), driven by the fact that reconstruction accuracy in the A condition fell below that of the V condition at −9 dB SNR ($t_{(20)} = 2$, $p =$

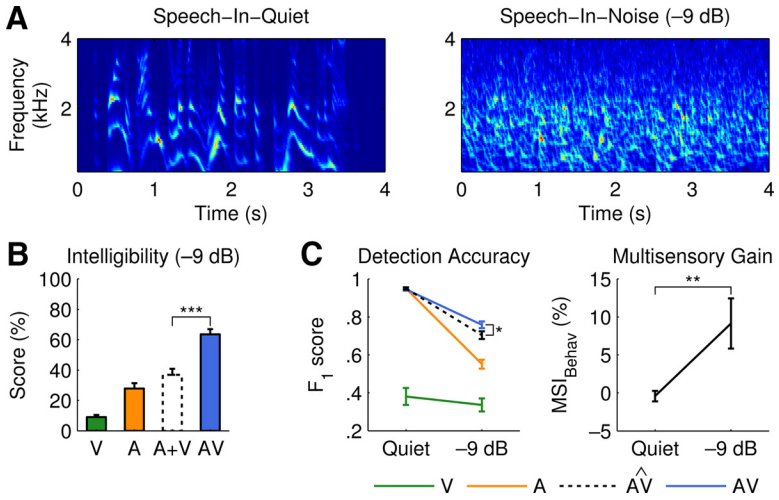

Figure 1. Audio stimuli and behavioral measures. ***A***, Spectrograms of a 4 s segment of speech-in-quiet (left) and speech-in-noise (−9 dB; right). ***B***, Subjectively rated intelligibility for speech-in-noise reported after each 60 s trial. White bar represents the sum of the unisensory scores. Error bars indicate SEM across subjects. Brackets indicate pairwise statistical comparisons (*$p < 0.05$; **$p < 0.01$; ***$p < 0.001$). ***C***, Detection accuracy (left) of target words represented as F_1 scores. The dashed black trace represents the statistical facilitation predicted by the unisensory scores. Multisensory gain (right) is represented as a percentage of unisensory performance.

0.055; A: 0.17 ± 0.05, V: 0.13 ± 0.04). Multisensory integration was indexed by differences in reconstruction accuracy between the AV condition and the A+V model. There was a main effect of condition across SNRs ($F_{(1,40)} = 115.1$, $p = 2.4 \times 10^{-13}$), with significantly higher reconstruction accuracy in the AV condition for both speech-in-quiet ($t_{(20)} = 7.1$, $p = 7.3 \times 10^{-7}$; AV: 0.2 ± 0.04, A+V: 0.18 ± 0.04) and speech-in-noise ($t_{(20)} = 8.1$, $p = 1 \times 10^{-7}$; AV: 0.16 ± 0.05, A+V: 0.14 ± 0.05). Although there was no significant interaction between SNR and condition ($F_{(1,40)} = 2.5$, $p = 0.12$), the multisensory gain (i.e., the AV enhancement as a percentage of A+V) was significantly greater at −9 dB SNR than in quiet [unpaired *t* test: $t_{(20)} = 2.8$, $p = 0.008$; MSI_{EEG} (Quiet): 10.6 ± 6.8%, MSI_{EEG} (−9 dB): 20.7 ± 14.9%; Fig. 2*A*, right]. These findings demonstrate that envelope tracking is restored in adverse hearing conditions by the addition of visual speech and that this process conforms to the principle of inverse effectiveness.

To examine the time lags that contributed most toward reconstruction, 33 separate estimates of the speech envelope were reconstructed using single-lag decoders between 0 and 500 ms (Fig. 2*B*). In all three conditions, the time lags that contributed the most information peaked at a later stage for speech-in-noise than for speech-in-quiet (Mann–Whitney *U* tests: $p < 0.05$). Running *t* tests comparing reconstruction accuracy in the AV condition with that of A+V at each time lag indicated that multisensory interactions occurred over a broad time window that was later for speech-in-noise than for speech-in-quiet ($p < 0.05$, Holm–Bonferroni corrected). It is likely that this difference in latency was primarily driven by the significant delay in envelope tracking observed in the A condition for speech-in-noise. Reconstruction accuracy in the A condition was also significantly lower than that of the V condition between 0 and 95 ms for speech-in-noise (running *t* test: $p < 0.05$, Holm–Bonferroni corrected). This suggests that in adverse hearing conditions, the sensitivity of auditory cortex to natural speech is significantly reduced during an early stage of the speech processing hierarchy.

Neural enhancement predicts behavioral gain

To investigate the relationship between our neural and behavioral measure of multisensory integration, we calculated the correlation coefficient between them using the reconstructed estimates from each of the 33 single-lag decoders. The logic here was that our behavioral multisensory effect may be reflected in our neural measure at a specific latency and integrating across 500 ms may obscure any correlation between these measures. Figure 2*C* shows the correlation between MSI_{Behav} and MSI_{EEG} at every time lag between 0 and 500 ms. There is no meaningful correlation for the first 200 ms, after which it begins to steadily increase until it peaks between 220 and 250 ms, at which latencies there is a significant (and-positive) correlation ($r = 0.44$, $p = 0.04$; Fig. 2*D*, left). This correlation is also significant if MSI is represented as percentage gain ($r = 0.56$, $p = 0.009$; Fig. 2*D*, right). If we calculate a linear fit to these data, the slope of the resulting line is ~0.96, meaning that, on average, a 50% gain in envelope tracking reflects a 52% gain in detection accuracy.

AV speech processing at multiple timescales

The timescale of AV speech processing has been closely linked to the rate at which syllables occur in extended passages of natural speech (Chandrasekaran et al., 2009; Luo et al., 2010; Crosse et al., 2015). To examine the impact of background noise on the timescale at which AV speech is integrated, we calculated the correlation coefficient between the reconstructed and original envelope at every 1 Hz frequency band between 1 and 30 Hz. Figure 3*A* shows the spectral profile of reconstruction accuracy for the AV condition and the A+V model. This spectrum represents the contribution of each frequency band to reconstructing the broadband envelope. Because the spectrum is consistently low pass in shape, we defined the cutoff frequency as the highest frequency at which reconstruction accuracy was greater than chance level (permutation test). For speech-in-quiet, reconstruction accuracy was greater than chance at frequencies between 1 and 8 Hz (Fig. 3*A*, left), whereas, for speech-in-noise, reconstruction accuracy was only greater than chance between 1 and 5 Hz (Fig. 3*A*, right).

Figure 3*B* shows the multisensory enhancement measured at each frequency band. To test for significance, paired *t* tests were conducted at only the frequencies at which reconstruction accuracy was greater than chance level ($p < 0.05$, Holm–Bonferroni corrected). For speech-in-quiet, there was a significant AV enhancement between 1 and 6 Hz (Fig. 3*B*, top), whereas for speech-in-noise, there was only a significant enhancement between 1 and 3 Hz (Fig. 3*B*, bottom). Although there were significant MSI effects across more frequency bands in quiet than in noise, it is important to note that this result does not contradict the principle of inverse effectiveness for the following reasons. First, performance values such as correlation coefficients should not be summed across frequency bands to arrive at a broadband measure of MSI. This can only be done using broadband speech itself. Second, we are using an absolute measure of MSI here

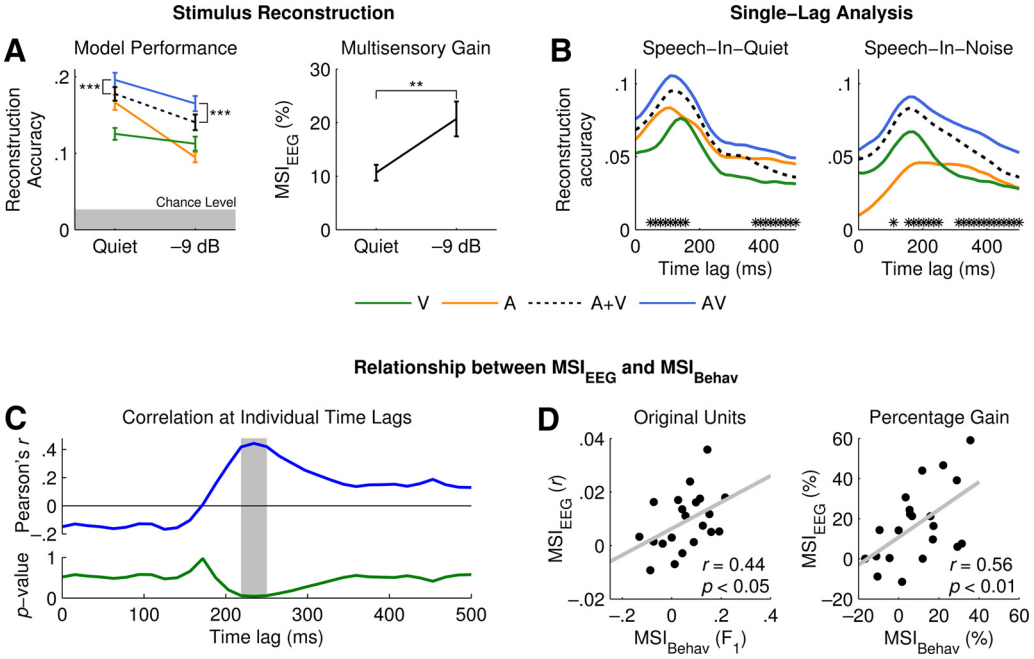

Figure 2. Stimulus reconstruction and relationship with behavior. ***A***, Reconstruction accuracy (left) obtained using decoders that integrated EEG across a 500 ms window. The dashed black trace represents the unisensory additive model. The shaded area indicates the 95th percentile of chance-level reconstruction accuracy (permutation test). Multisensory gain (right) represented as a percentage of unisensory performance. Error bars indicate SEM across subjects. Brackets indicate pairwise statistical comparisons (**$p < 0.01$; ***$p < 0.001$). ***B***, Reconstruction accuracy obtained using single-lag decoders at every lag between 0 and 500 ms. The markers running along the bottom of each plot indicate the time lags at which MSI_{EEG} is significant ($p < 0.05$, Holm–Bonferroni corrected). ***C***, Correlation coefficient (top) and corresponding *p*-value (bottom) between MSI_{EEG} and MSI_{Behav} at individual time lags for speech-in-noise. The shaded area indicates the lags at which the correlation is significant or trending toward significance (220–250 ms; $p < 0.05$). ***D***, Correlation corresponding to shaded area in ***C*** with MSI_{EEG} and MSI_{Behav} represented in their original units (left) and as percentage gain (right).

because we are comparing it with no MSI (i.e., zero). Because we are not using a relative measure of MSI (i.e., multisensory gain), we therefore cannot compare MSI values between listening conditions directly at each frequency band.

To relate these findings to the temporal scale of natural speech, we summarized the average rate of different linguistic units by deriving the durations of the respective speech segments from the audio files (Fig. 3*C*). The results suggest that, in quiet, AV speech was integrated at frequencies commensurate with the rate of suprasegmental information such as sentential and phrasal units, as well as smaller segmental units such as words and syllables. In background noise, AV integration was only evident at the sentential and lexical timescale.

AV temporal integration

Given that background-insensitive speech recognition has been linked to long-term temporal integration (Ding and Simon, 2013), we wished to examine the role of temporal integration in maintaining AV speech processing in background noise. The decoder window size was shortened from 500 to 100 ms in steps of 100 ms, restricting the amount of temporal information that each decoder could integrate across when reconstructing the stimulus. Although this reduced decoder performance in both quiet (ΔAV: 0.04 ± 0.01) and in noise (ΔAV: 0.06 ± 0.03), the effect was significantly greater in the latter (unpaired *t* test: $t_{(40)} = 2.7$, $p = 0.01$; Fig. 4*A*). As a result, multisensory gain was more sensitive to modulations in temporal window size in noise ($F_{(1.8,36.5)} = 1.4$, $p = 0.27$, one-way ANOVA) than in quiet ($F_{(1.3,26.7)} = 0.31$, $p = 0.87$, one-way ANOVA). Although the effect was not significant, MSI_{EEG} decreased as the temporal window size was reduced (Fig. 4*B*). Critically, inverse effectiveness (as indexed by the difference between MSI_{EEG} in quiet and at −9 dB) was only significantly greater than zero when the decoders integrated EEG over temporal window sizes of >300 ms (unpaired *t* tests: $p < 0.05$; Fig. 4*C*).

Discussion

Our findings exhibit three major electrophysiological features of AV speech processing. First, the accuracy with which cortical activity entrains to AV speech conforms to the principle of inverse effectiveness. Second, visual speech input restores early tracking of the acoustic speech signal in background noise and is integrated with auditory information at much lower frequencies. Third, inverse effectiveness in natural AV speech processing relies on crossmodal integration over long temporal windows. Our findings suggest that AV speech integration is maintained in background noise by several underlying mechanisms.

Envelope tracking and inverse effectiveness

Consistent with seminal work on AV speech-in-noise (Sumby and Pollack, 1954; Ross et al., 2007), we demonstrated that the behavioral benefit produced by AV speech was significantly greater in noise than in quiet. This inverse effectiveness phenomenon was also observed in our EEG data, which revealed that multisensory interactions were contributing to the neural tracking of AV speech to a greater extent in noise than in quiet. In support of our neuronal effect, a recent MEG study demonstrated (using a phase-based measure of neural tracking) that coherence across multiple neural response trials was enhanced by AV speech relative to A speech when participants listened to competing speakers, but not single speakers (Zion Golumbic et al., 2013). In other words, making it more difficult to hear the target speaker by introducing a second speaker revealed an enhancement in AV speech tracking that was not detectable in single-speaker speech.

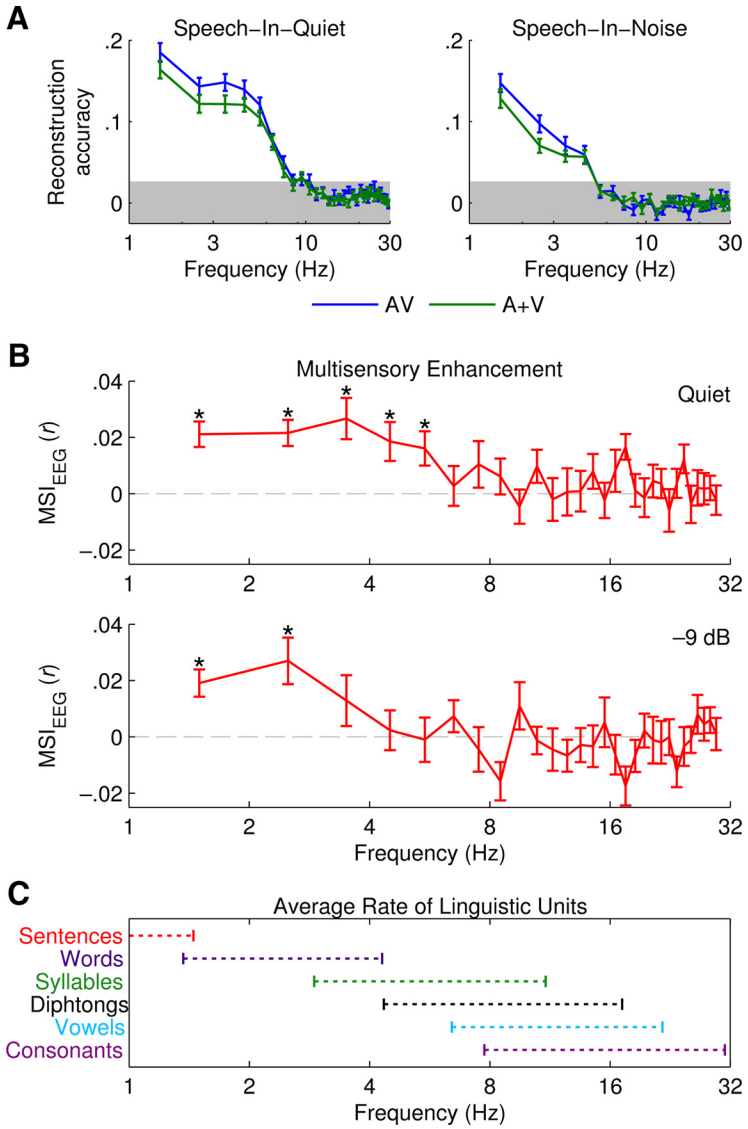

Figure 3. AV speech integration at multiple timescales. ***A***, Reconstruction accuracy for AV (blue) and A+V (green) at each frequency band. The shaded area indicates the 5th to 95th percentile of chance-level reconstruction accuracy (permutation test). Error bars indicate SEM across subjects. ***B***, Multisensory enhancement at each frequency band. The markers indicate frequency bands at which there was a significant multisensory interaction effect ($p < 0.05$, Holm–Bonferroni corrected). ***C***, Average rate of different linguistic units derived from the audio files of the speech stimuli using phoneme-alignment software. The brackets indicate mean ± SD.

For speech-in-noise, we found that the multisensory enhancement in envelope tracking at 220–250 ms accurately predicted the multisensory gain in behavior. To interpret the significance of this temporal locus, we must first consider what each of these multisensory indices reflect. Our behavioral measure (MSI_{Behav}) was derived from the accuracy with which participants detected target words. Because the task involved identifying whole words, the MSI score may reflect cross-modal integration at the semantic level (Ross et al., 2007). In support of this, the time course of speech perception in the superior temporal cortex has been shown to reflect lexical–semantic processing from 200 ms onwards (Salmelin, 2007; Picton, 2013). Our neural measure (MSI_{EEG}) was derived from how accurately the speech envelope could be reconstructed from the EEG data. Specifically, we observed multisensory interactions below 3 Hz in noise over a broad range of time lags. Given that this frequency range is commensurate with the average rate of spoken words, it fits well with our behavioral task. Furthermore, neural oscillations in the delta range (1–4 Hz) are thought to integrate cross-modal information over a temporal window of ~125–250 ms (Schroeder et al., 2008), consistent with our broad window of integration. It is likely that this broad window reflects neural integration at multiple stages of the speech-processing hierarchy. However, given that our behavioral measure of multisensory integration likely reflects processing at a more specific (lexical–semantic) stage of processing, the correlation that we saw between behavioral and neural integration was only evident at a latency that relates to this stage of the speech-processing hierarchy.

AV mechanisms in speech-in-noise

Our EEG data suggest that cortical activity entrains to AV speech only at lower frequencies in background noise. In support of this notion, it has been demonstrated that MEG entrains to AV speech at lower frequencies when a competing speaker is introduced (Zion Golumbic et al., 2013). An MEG study by Ding and Simon (2013) that investigated neural entrainment to audio-only speech at different SNRs found that the cutoff frequency of the phase-locking spectrum decreased linearly with SNR, but that low delta-band neural entrainment was relatively insensitive to background noise above a certain threshold. This mechanism of contrast gain control was linked to the M100 component of the temporal response function (TRF), which was shown to be relatively robust to noise, unlike the earlier M50 component (Ding and Simon, 2013; Ding et al., 2014). Our results, along with these other studies, indicate that low-frequency speech information is more reliably encoded than higher-frequency linguistic content in adverse hearing conditions and that this process is likely maintained by contrast gain control and adaptive temporal sensitivity in auditory cortex (Ding and Simon, 2013).

In addition, we found that auditory and visual information interacted at lower frequencies in noise than in quiet, which is unsurprising given that there is a more robust auditory representation encoded at lower frequencies. Consistent with this, we showed that inverse effectiveness relied on longer temporal windows of integration, something that is also critical for a noise-robust cortical representation of speech (Ding and Simon, 2013). A recent intracranial study that examined AV integration in quiet using discrete, nonspeech stimuli, observed multisensory enhancement effects [AV − (A+V)] in delta- and theta-phase alignment (Mercier et al., 2015). Interestingly, they reported visually driven crossmodal delta-band phase-reset in auditory cortex. It is possible that this process could be mediated by delta-frequency head movements, which have been shown to convey

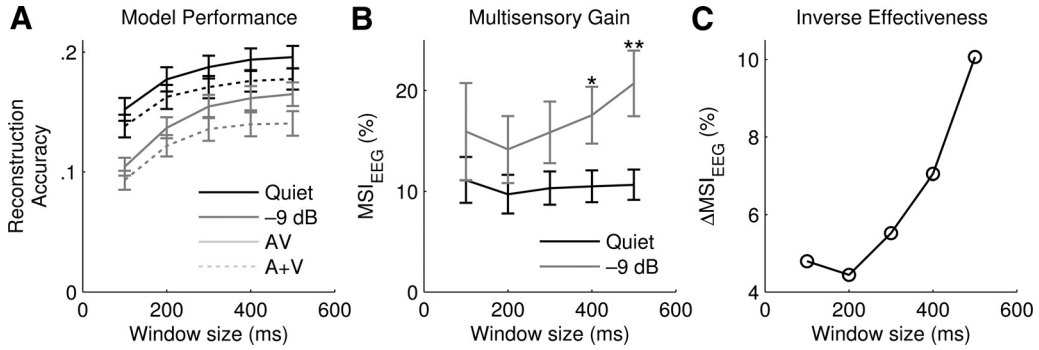

Figure 4. AV temporal integration. ***A***, Model performance by decoder temporal window size. Error bars indicate SEM across participants. ***B***, Multisensory gain by decoder temporal window size. Markers indicate window sizes at which there was significant inverse effectiveness (i.e., −9 dB > quiet; $*p < 0.05$; $**p < 0.01$). ***C***, Inverse effectiveness by decoder temporal window size.

prosodic information important to speech intelligibility (Munhall et al., 2004). Integration of auditory and visual speech information could be maintained in adverse hearing conditions by a combination of delta-frequency phase resetting and long-term temporal integration.

Multistage integration model

As mentioned earlier, a growing body of evidence indicates that multisensory integration likely occurs over multiple temporal stages during AV speech processing (Schwartz et al., 2004; van Wassenhove et al., 2005; Eskelund et al., 2011; Baart et al., 2014; Peelle and Sommers, 2015). The findings presented here will be interpreted within the context of such multistage integration models and, in particular, the role of prediction and constraint as early and late integration mechanisms, respectively (Peelle and Sommers, 2015).

The notion that an early integration mechanism increases auditory cortical sensitivity seems highly relevant in the context of speech-in-noise. Here, we demonstrated that neural tracking of audio-only speech in noise was significantly diminished at time lags between 0 and 95 ms, suggesting that auditory cortical sensitivity was reduced at an early stage of speech processing. Although the current data indicate that envelope tracking was restored by the addition of visual speech input at this early processing stage, because we include the entire head during the reconstruction analysis, it is difficult to say whether this is the result of increased auditory cortical sensitivity or rather contributions from multisensory areas such as STS or visual cortical areas. Furthermore, our single-lag analysis did not reveal significant crossmodal interactions at this early stage. However, a theory that supports this notion of an early increase in auditory cortical sensitivity is that of cross-sensory phase-resetting of auditory cortex (Lakatos et al., 2007; Kayser et al., 2008; Schroeder et al., 2008; Arnal et al., 2009; Mercier et al., 2015). Although such a mechanism can be difficult to reconcile in the context of extended vocalizations given that the time lag between visual and auditory speech is so variable (Schwartz and Savariaux, 2014), this can be explained somewhat by the temporal correspondence between the hierarchical organization of speech and that of the rhythmic oscillations in primary auditory cortex (Schroeder et al., 2008; Giraud and Poeppel, 2012). Although, intuitively, it may seem more likely that auditory cortex would be primed by continuous visual input in a tonic manner, the idea of phasic crossmodal priming is supported by the fact that the temporal coherence between the A and V streams is critical for enhanced neural tracking during AV speech (Crosse et al., 2015). This is also supported by accounts of enhanced phasic coordination across auditory and visual cortices for matched versus mismatched AV stimuli (Luo et al., 2010).

Evidence of a later integration stage that constrains lexical selection can also be found in numerous electrophysiological studies. Both TRF and event-related potential measures have revealed emergent multisensory interaction effects in the form of a reduced component amplitude (Besle et al., 2004; van Wassenhove et al., 2005; Bernstein et al., 2008; Crosse et al., 2015). This reduction in cortical activation may well reflect a mechanism that constrains lexical computations based on the content of preceding visual information. Both our single-lag analysis and temporal window analysis further suggest that integrating later temporal information contributes to AV speech processing. However, the most compelling evidence that is provided in favor of a late integration stage is the correspondence that was observed between the behavioral and neural measures at 220–250 ms. Given the likelihood that both of these measures reflect integration at the lexical–semantic level fits well with current views on the time course of the auditory processing hierarchy (Salmelin, 2007; Picton, 2013).

In summary, our results support the theory that visual speech input restores early tracking of auditory speech and subsequently constrains lexical processing at a later computational stage. We contend that inverse effectiveness, in the context of AV speech processing, relies heavily on our ability to integrate crossmodal information over longer temporal windows in background noise.

References

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

Baart M, Stekelenburg JJ, Vroomen J (2014) Electrophysiological evidence for speech-specific audiovisual integration. Neuropsychologia 53:115–121. CrossRef Medline

Bernstein LE, Auer ET Jr, Takayanagi S (2004) Auditory speech detection in noise enhanced by lipreading. Speech Communication 44:5–18. CrossRef

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

Blamey PJ, Cowan RS, Alcantara JI, Whitford LA, Clark GM (1989) Speech perception using combinations of auditory, visual, and tactile information. J Rehabil Res Dev 26:15–24. Medline

Campbell R (2008) The processing of audio-visual speech: empirical and neural bases. Philos Trans R Soc 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

Crosse MJ, Butler JS, Lalor EC (2015) Congruent visual speech enhances cortical entrainment to continuous auditory speech in noise-free conditions. J Neurosci 35:14195–14204. CrossRef Medline

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

Ding N, Simon JZ (2012) Neural coding of continuous speech in auditory cortex during monaural and dichotic listening. J Neurophysiol 107: 78–89. 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, 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

Erber NP (1975) Auditory-visual perception of speech. J Speech Hear Disord 40:481–492. CrossRef Medline

Eskelund K, Tuomainen J, Andersen TS (2011) Multistage audiovisual integration of speech: dissociating identification and detection. Exp Brain Res 208:447–457. 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 (2001) The effect of speechreading on masked detection thresholds for filtered speech. J Acoust Soc Am 109:2272–2275.

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

Irino T, Patterson RD (2006) A dynamic compressive gammachirp auditory filterbank. IEEE Trans Audio Speech Lang Process 14:2222–2232. CrossRef Medline

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

Lakatos P, Chen CM, O'Connell MN, Mills A, Schroeder CE (2007) Neuronal oscillations and multisensory interaction in primary auditory cortex. Neuron 53:279–292. CrossRef Medline

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

Ma WJ, Zhou X, Ross LA, Foxe JJ, Parra LC (2009) Lip-reading aids word recognition most in moderate noise: a Bayesian explanation using high-dimensional feature space. PLoS One 4:e4638.

Mercier MR, Molholm S, Fiebelkorn IC, Butler JS, Schwartz TH, Foxe JJ (2015) Neuro-oscillatory phase alignment drives speeded multisensory response times: an electro-corticographic investigation. J Neurosci 35: 8546–8557. CrossRef Medline

Meredith MA, Stein BE (1986) Spatial factors determine the activity of multisensory neurons in cat superior colliculus. Brain Res 365:350–354. 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

Munhall KG, Jones JA, Callan DE, Kuratate T, Vatikiotis-Bateson E (2004) Visual prosody and speech intelligibility head movement improves auditory speech perception. Psychol Sci 15:133–137. CrossRef Medline

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

Parsons TW (1987) Voice and speech processing. New York: McGraw-Hill College.

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

Picton T (2013) Hearing in time: evoked potential studies of temporal processing. Ear Hear 34:385–401. CrossRef Medline

Van Rijsbergen CJ (1979) Information retrieval. London: Butterworth-Heinemann.

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

Salmelin R (2007) Clinical neurophysiology of language: The MEG approach. Clin Neurophysiol 118:237–254. 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

Schwartz JL, Savariaux C (2014) No, there is no 150 ms lead of visual speech on auditory speech, but a range of audiovisual asynchronies varying from small audio lead to large audio lag. PLoS Comput Biol 10:e1003743. CrossRef Medline

Schwartz JL, Berthommier F, Savariaux C (2004) Seeing to hear better: evidence for early audio-visual interactions in speech identification. Cognition 93:B69–B78. CrossRef Medline

Stevenson RA, Ghose D, Fister JK, Sarko DK, Altieri NA, Nidiffer AR, Kurela LR, Siemann JK, James TW, Wallace MT (2014) Identifying and quantifying multisensory integration: a tutorial review. Brain Topogr 27: 707–730. 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. London: Lawrence Erlbaum Associates.

Theunissen FE, Sen K, Doupe AJ (2000) Spectral-temporal receptive fields of nonlinear auditory neurons obtained using natural sounds. J Neurosci 20:2315–2331. 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

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