Behavioral/Cognitive

# Visual Input Enhances Selective Speech Envelope Tracking in Auditory Cortex at a "Cocktail Party"

**Elana Zion Golumbic,**[1,2] **Gregory B. Cogan,**[3] **Charles E. Schroeder,**[1,2] **and David Poeppel**[3]

[1]Department of Psychiatry, Columbia University College of Physicians and Surgeons, New York, New York, 10032; [2]Cognitive Neuroscience and Schizophrenia Program, Nathan S. Kline Institute for Psychiatric Research, Orangeburg, New York, 10962; [3]Department of Psychology, New York University, New York, 10003

**Our ability to selectively attend to one auditory signal amid competing input streams, epitomized by the "Cocktail Party" problem, continues to stimulate research from various approaches. How this demanding perceptual feat is achieved from a neural systems perspective remains unclear and controversial. It is well established that neural responses to attended stimuli are enhanced compared with responses to ignored ones, but responses to ignored stimuli are nonetheless highly significant, leading to interference in performance. We investigated whether congruent visual input of an attended speaker enhances cortical selectivity in auditory cortex, leading to diminished representation of ignored stimuli. We recorded magnetoencephalographic signals from human participants as they attended to segments of natural continuous speech. Using two complementary methods of quantifying the neural response to speech, we found that viewing a speaker's face enhances the capacity of auditory cortex to track the temporal speech envelope of that speaker. This mechanism was most effective in a Cocktail Party setting, promoting preferential tracking of the attended speaker, whereas without visual input no significant attentional modulation was observed.**

**These neurophysiological results underscore the importance of visual input in resolving perceptual ambiguity in a noisy environment. Since visual cues in speech precede the associated auditory signals, they likely serve a predictive role in facilitating auditory processing of speech, perhaps by directing attentional resources to appropriate points in time when to-be-attended acoustic input is expected to arrive.**

## Introduction

Understanding speech, particularly under noisy conditions, is significantly facilitated by viewing the speaker's face (Sumby and Pollack, 1954; Grant and Seitz, 2000; Schwartz et al., 2004). Multiple studies indicate that visual input affects neural responses to speech, both in early sensory cortices and higher order speech-related areas (Besle et al., 2004; Davis et al., 2008; McGettigan et al., 2012). However, little is known about the neural dynamics by which visual input facilitates on-line speech perception, since the majority of electrophysiological studies to date have focused on audiovisual (AV) effects of processing individual syllables, in rather unnatural laboratory paradigms (van Wassenhove et al., 2005).

Recently, important advances have been made in the ability to quantify the neural response to continuous speech. Converging evidence across several methodologies indicates that low-frequency neural activity in auditory cortex (<15 Hz) phase lock to the temporal envelope of speech, which fluctuates at similar rates (Rosen, 1992; Luo and Poeppel, 2007; Aiken and Picton, 2008; Lalor and Foxe, 2010; Hertrich et al., 2012; Peelle et al., 2012). The temporal envelope of speech is critical for comprehension (Shannon et al., 1995; Drullman, 2006) and it has been suggested that this "envelope-tracking" response serves to parse the continuous input into smaller units (syllables or phrases) to which higher order decoding of the fine structure is applied, allowing for syllable classification and word recognition (Ghitza, 2011; Giraud and Poeppel, 2012).

To date, studies investigating the envelope-tracking response have mainly used auditory stimuli. However, articulatory facial movements are also correlated with the speech envelope and precede it by ~150 ms (Grant and Seitz, 2000; Kim and Davis, 2003; Chandrasekaran et al., 2009). Thus, theoretically, viewing the talking face can provide predictive cues for upcoming speech and facilitate its processing (van Wassenhove et al., 2005; Schroeder et al. 2008; Arnal et al. 2011). In this paper we investigate whether viewing the speaker's face indeed enhances the envelope-tracking response in auditory cortex.

Since the benefit of congruent visual input to speech perception is greatest under difficult auditory conditions (Sumby and Pollack, 1954; Callan et al., 2003; Ross et al., 2007), we investigated AV effects on the envelope-tracking response under two conditions: when listening to a single speaker and when attending to one speaker while ignoring a concurrent irrelevant speaker (Cherry, 1953; McDermott, 2009 (simulating a "Cocktail Party"). Previous studies, using auditory-only stimuli, have shown that the envelope-tracking response of the attended speaker is ampli-

Received Aug. 1, 2012; revised Oct. 17, 2012; accepted Oct. 30, 2012.

Author contributions: E.Z.G., G.B.C., C.E.S., and D.P. designed research; E.Z.G. and G.B.C. performed research; E.Z.G. and G.B.C. analyzed data; E.Z.G., G.B.C., C.E.S., and D.P. wrote the paper.

This work was supported by funding from the following grants: National Institutes of Health 2R01DC 05660 to D.P., P50 MH086385 to C.E.S., and F32 MH093601 to E.Z.G.

Correspondence should be addressed to Elana Zion Golumbic, Division for Therapeutic Brain Stimulation, Department of Psychiatry, Columbia University College of Physicians and Surgeons, New York, NY 10032. E-mail: ezg2101@columbia.edu.

DOI:10.1523/JNEUROSCI.3675-12.2013

fied compared with the ignored speaker (Kerlin et al., 2010; Ding and Simon, 2012; Mesgarani and Chang, 2012). Nonetheless, the response to the ignored speaker remains robust, which can lead to behavioral interference (Cherry, 1953; Moray, 1959; Wood and Cowan, 1995; Beaman et al., 2007).

Here we recorded magnetoencephalographic (MEG) signals as human participants attended to segments of natural speech, presented either with or without the corresponding talking face. We investigated whether viewing the talking face enhances the envelope-tracking response and whether visual input improves the preferential tracking of the attended speaker in a Cocktail Party environment (Zion Golumbic et al., 2012).

## Materials and Methods

### *Participants*

Thirteen native English-speaking participants (eight female, median age 22, one left handed) with normal hearing and no history of neurological disorders provided informed consent according to the University Committee on Activities Involving Human Participants at New York University. All participants but one were right handed as assessed by the Edinburgh Inventory of Handedness (Oldfield, 1971).

### *MEG recordings*

MEG data were collected on a 157-channel whole-head MEG system (5 cm baseline axial gradiometer SQUID-based sensors; KIT) in an actively magnetically shielded room (Vakuumschmelze GmbH). Data were sampled at 1000 Hz, with a notch filter at 60 Hz, and an on-line recording 200 Hz lowpass filter. Each participant's head position was assessed via five coils attached to anatomical landmarks both before and after the experiment to ensure that head movement was minimal. Head-shape data were digitized using a 3D digitizer (Polhemus). The auditory signals were presented though in-ear earphones (Etymotic ER3-A) and the speech sounds were presented at comfortable conversational levels (~72 dB SPL). The visual materials were presented on a rear-projection screen in the shielded room (~18° horizontal and 11° vertical visual angles, ~ 44 cm from eyes; Infocus LP 850 projector). Stimulus delivery and triggering were controlled by Presentation (Neurobehavioral Systems).

### *Experimental design*

The stimuli consisted of eight movie clips of two speakers (one male, one female; four movies per speaker) reciting a short passage (9.8 ± 1.5 s). The movies were edited using QuickTime Pro (Apple) to align the faces in the center of the frame, equate the relative size of the male and female faces, and to clip the movies appropriately. Each female movie was paired with a male movie of approximately similar length (<0.5 s difference in length), and this pairing remained constant throughout the entire study, yielding four stimulus pairs. In each trial, a stimulus pair was selected, and either one of the stimuli from the pair was presented individually (Single Speaker) or both stimuli were presented simultaneously (Cocktail Party). The stimuli were presented either with or without the video (AV/A), yielding a total of four conditions: AVsingle, AVcocktail, Asingle, Acocktail (see Fig. 1*A*). To ensure sufficient repetitions of each stimulus in a particular same condition, the attribution of stimulus pair to condition was held constant across the experiment. Specifically, two stimulus pairs were used for the AV conditions and two for the A conditions, and in both cases the same stimulus pairs were used for the Single Speaker and Cocktail Party conditions. The envelopes of all stimulus pairs were uncorrelated (Pearson correlation coefficient $r$ <0.065 for all pairs).

In all conditions, the audio signal was presented diotically, so the auditory streams could not be segregated based on spatial cues. The videos were presented on either side of a computer screen, and the location of each speaker (left/right) was assigned randomly in each trial. In the A conditions, rectangular placeholders were presented on both sides of the screen instead of the videos.

Before each trial, instructions appeared in the center of the screen indicating which of the speakers to attend to (e.g., "Attend Female"). The participants indicated with a button press when they were ready to begin, and the stimuli started playing 2 s after their response. The location of the to-be-attended speaker was highlighted by a red frame and the verbal instruction remained in the center of the screen, to ensure the participants remembered which speaker to attend to during the entire trial. Each stimulus was cut off before the speaker uttered the last word, and then a target word appeared in the center of the screen. Participants' explicit task was to indicate via button press whether the target word was a congruent ending to the attended passage. For example: passage: "…my parents thought that idea was" Target words: silly/amusing/funny (congruent); purple/cloudy/hanging (incongruent). Target words were unique on each trial (no repetitions), and 50% were congruent with the attended segment. Progression to the next trial was self-paced.

There were a total of 40 trials in each condition, with each individual stimulus designated as the "attended" stimulus in 10 trials. The order of the trials was randomized throughout the experiment. Breaks were given approximately every 10 min, and the total duration of the experiment was ~45 min.

### *Data analysis*

All analyses were performed using MATLAB (MathWorks) and the Fieldtrip toolbox (Oostenveld et al., 2011). The data were noise reduced off-line using a time-shift Principled Component Analysis (de Cheveigné and Simon, 2007). Ocular and cardiac artifacts were corrected using ICA decompositions (Jung et al., 2000). The data were visually inspected and trials with large artifacts were removed from further analysis. The data were initially segmented into trials starting 4 s before stimulus onset and lasting for 16 s poststimulus to include the entire duration of all stimuli and to avoid contamination of edge effects in subsequent filtering and spectral analysis.

Although behavioral performance was ~80% correct, we nonetheless included both correct and incorrect trials in our analysis. This decision was mainly due to the small number of trials per stimulus ($n = 10$), since the electrophysiological measures used here are highly sensitive to number of trials and need to be equated across stimuli. Due to limitations on the total length of the experiment it was not feasible to substantially increase the number of trials. We recognize that this may potentially weaken the size of our effects, since we cannot know if incorrect responses are due to lapses of attention or whether attention was appropriately allocated, but participants made mistakes in the comprehension task. Nonetheless, we argue that any significant effects found despite this caveat are valid, because including incorrect trials works against our hypothesis.

We performed two complementary analyses to evaluate the envelope-tracking responses in the MEG signal.

*Phase dissimilarity/selectivity analysis.* For the Single Speaker trials, we computed a "phase-dissimilarity index," introduced by Luo and Poeppel (2007), which characterizes the consistency and uniqueness of the temporal (phase) pattern of neural responses to different speech tokens. The rationale behind this analysis is to compare the phase consistency across repetitions of the same stimulus (within-stimulus) with a baseline of phase consistency across trials in which different stimuli were presented (across-stimuli).

Since the stimuli differed somewhat in their duration, this analysis focused on the first 8 s of each epoch, matching the duration of the shortest stimulus. For each participant and each sensor, we first estimated the momentary phase in single trials. Since previous studies have shown phase dissimilarity effects primarily in frequencies <10 Hz, we performed a wavelet decomposition of single trials using a complex 6-cycle Morlet wavelet in logarithmic steps between 0.5 and 15 Hz, resulting in 51 frequency points. Next, we calculated the intertrial phase-locking value (ITC; Eq. 1) at each time-frequency point, across all trials in which the same stimulus was presented. For each frequency level, the ITC time course was averaged over time (0–8 s) and across all stimuli to obtain the average within-stimulus ITC.

$$ITC = \left| \sum_{j=1}^{N} e^{i\varphi_j(t,f)} \right| \tag{1}$$

The across-stimuli ITC was estimated using the same approach but using shuffled data, such that that the ITC was computed across randomly selected trials in which different stimuli were presented. The phase dissimilarity index is computed as the difference between the within-stimuli and the across-stimuli ITC (Eq. 2a). Large phase-dissimilarity values indicate that the responses to individual stimuli have a highly consistent time course as evidenced in the response to single trials. This analysis was performed separately for the A and AV trials.

$$(a): phase\_dissimilarity(f) = ITC(f)_{within\_stim} - ITC(f)_{across\_stim}$$

$$(b): phase\_selectivity(f) = ITC(f)_{within\_attention} - ITC(f)_{across\_attention} \quad (2)$$

For the Cocktail Party trials, we computed a "phase-selectivity index," which is based on the same logic as the phase-dissimilarity index but is designed to determine how attention modulates the temporal pattern of the neural response, given a particular pair of speakers. To this end, we calculated the within-attention and across-attention ITC (Eq. 2b), defined as follows. The within-attention ITC was computed across all trials in which the same pair of speakers was presented and the same speaker was attended. The across-attention ITC was computed across trials in which the same pair of speakers was presented, but with a random mixture of attend-female and attend-male trials. The within-attention and across-attention ITCs are then averaged over time and stimulus pairs and subtracted from each other yielding the phase-selectivity index (Eq. 2b).

Large phase-selectivity values indicate that the time course of the neural responses is influenced by attention and is substantially different when attending to different stimuli, despite the identical acoustic input. In contrast, low phase selectivity would suggest similar patterns of the neural responses despite attending to different speakers, a pattern that likely represents a mixture of responses to both speakers, which is not modulated by attention.

To select channels for statistical analysis, we collapsed the phase-dissimilarity and phase-selectivity indices across all conditions, frequencies, and participants, and selected the 20 channels with the highest averaged values. The procedure ensured that channel selection was not biased by condition or frequency band. We then averaged the phase-dissimilarity and phase-selectivity indices across those 20 channels separately for each condition and frequency band. Since ITC values are not normally distributed, we applied a rau transformation phase-dissimilarity and phase-selectivity indices before make them suitable for linear parametric statistical testing (Studebaker, 1985). For each condition, we determined which frequencies had significant phase dissimilarity/selectivity using a $t$ test at each frequency level. We controlled for multiple comparisons by requiring clusters of at least four consecutive frequency points at a significance level of $p < 0.01$. To evaluate AV effects, we performed a paired $t$ test between the average phase dissimilarity/selectivity in the AV and A conditions, and separately for the Single Speaker and Cocktail Party conditions.

*Temporal response function*. To determine the relationship between the neural response and the presented speech stimuli, we estimated a linear temporal response function (TRF) between the stimulus and the response. The neural response $r(t)$ is modeled by the temporal envelope of the presented speaker $s(t)$ as follows:

$$r(t) = \sum_{\tau} s(t - \tau)\, TRF(\tau) + \varepsilon(t) \quad (3)$$

The $TRF(t)$ is a linear kernel and $\varepsilon(t)$ is the residual response not explained by the model (Ding and Simon, 2012). The broadband envelope of speech $s(t)$ was extracted by filtering the speech stimuli between 250 and 4000 Hz and extracting the temporal envelope using a Hilbert transform. The temporal response functions $TRF(t)$ were fitted using normalized reverse correlation as implemented in the STRFpak MATLAB toolbox (http://strfpak.berkeley.edu/) (Theunissen et al., 2001; Lalor and Foxe, 2010). Normalized reverse correlation involves inverting the autocorrelation matrix of the stimulus, which is usually numerically ill conditioned. Therefore, a pseudo-inverse is applied instead, which ignores eigenvalues of the autocorrelation matrix that are smaller than a predefined tolerance factor. The tolerance factor was scanned and determined by a pre-analysis to optimize the predictive power and then fixed for all sensors and participants.

TRF was estimated independently for each participant at each sensor, and separately for each of the four conditions. For the Single Speaker conditions, $r(t)$ was a concatenated vector of the responses to each stimulus averaged over trials, and $s(t)$ was a vector of the envelopes, concatenated in the same manner. In the Cocktail Party conditions, $r(t)$ was a concatenated vector of the responses to each attended-stimulus averaged over trials, and it was modeled by the temporal envelopes of both the attended and ignored speakers ($s_A(t)$ and $s_I(t)$, respectively), generating a temporal response function for each speaker ($TRF_A$ and $TRF_I$, respectively).

$$r(t) = \sum_{\tau} s_{\mathrm{A}}(t - \tau)\, TRF_{\mathrm{A}}(\tau) + \sum_{\tau} s_{\mathrm{I}}(t - \tau)\, TRF_{\mathrm{I}}(\tau) + \varepsilon(t) \quad (4)$$

If the two films presented in the same trial had different lengths, only the portion of the stimulus that overlapped in time was included in the model, and the response $r(t)$ to that stimulus pair was truncated accordingly. Both $r(t)$ and $s(t)$ were downsampled to 100 Hz before model estimation.

The TRFs were 300 ms long (30 estimated points) and were estimated using a jackknife cross-validation procedure to minimize effects of overfitting (Ding and Simon, 2012). In this procedure, given a total of $N$ stimuli, a TRF is estimated between $s(t)$ and $r(t)$ derived from $N - 1$ stimuli, and this estimate is used to predict the neural response to the left-out stimulus. The goodness of fit of the model was evaluated by the correlation between the actual neural response and the model prediction, called predictive power (David et al., 2007). The predictive power calculated from each jackknife estimate is averaged.

To evaluate the significance of the TRF estimate, we repeated the cross-validation procedure for each participants and each sensor on surrogate data, mismatching the stimulus and responses vectors. The statistical significance of the predictive power of TRF estimation from the real data was evaluated by comparing it to the predictive power of the surrogate TRFs using a paired $t$ test. Similarly, we evaluated the significance of the peak amplitude of the estimated TRF by comparing it to the amplitude of the surrogate TRFs at the same time point using a paired $t$ test.

### *Source reconstruction*

The analyses were repeated in source space for five of the participants for whom structural magnetic resonance imaging (MRI) was available. All source reconstructions were done using minimum norm estimation (Hämäläinen and Ilmoniemi, 1994).

Each participant's structural MRI was reconstructed using the FreeSurfer suite (http://surfer.nmr.mgh.harvard.edu/) to produce a 3D image of their MRI. This was used to localize neural activity onto the brain. The source space was set up such that each participant's brain surface contained ~20480 "triangles" of localized dipoles. Due to computational constraints, this value was downsampled using a triangulation procedure that recursively subdivides the inflated spherical surface into icosahedrons and then subdivides the number of triangles (sides) of these icosahedrons by a factor of four for the TRF analysis, and by a factor of 16 for the phase-dissimilarity calculations. This produced, for the TRF analysis, a source space with 2562 sources per hemisphere, with an average source spacing of ~6.2 mm. For the ITC calculations, this produced 642 sources per hemisphere, with an average source spacing of ~10 mm.

The forward solution was computed using the decimated reconstruction surface as well as the boundary element model information computed for a single compartment (homogenous) model for MEG data only.

The inverse operator was then computed using the forward solution as well as the noise covariance matrix computed from each participant; no task data were collected at the beginning of the experimental session. A depth weighting of 0.8 and a regularization parameter, $\lambda^2$, of 0.1 were used.

The orientation of the sources was fixed to be normal to cortical surface, as the primary source of the MEG signal is thought to origi-

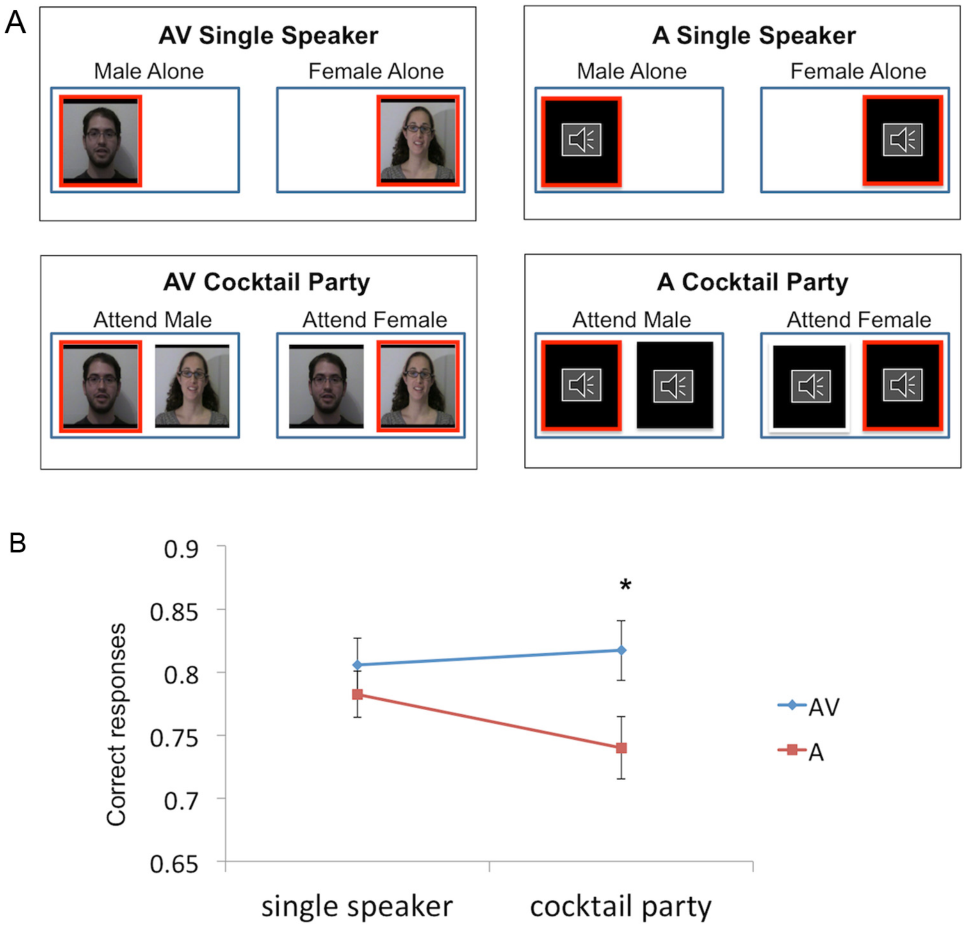

**Figure 1.** Paradigm and behavioral results. ***A***, Illustration of the four experimental conditions. In each trial, a stimulus pair was presented with or without the corresponding video of the speakers (AV/A) and either one speaker or both speakers were audible (Single Speaker/Cocktail Party). The red rectangle indicated which speaker was designated to be attended. Trial order was randomized throughout the experiment. ***B***, Ratio of correct responses, averaged across participants (±1 SEM). Performance was generally good in all conditions, indicating that participants were indeed attending to the prescribed speaker. Performance was reduced slightly, but significantly, in the A Cocktail Party condition suggesting that selective attention was more challenging in the absence of visual input.

nate from postsynaptic potentials of the apical dendrites of large pyramidal cells orientated perpendicular to the cortical surface (Hämäläinen and Ilmoniemi, 1994). Input data for the ITC calculations were the raw time series of the neural responses to each of the stimuli for each participant. Single trial time-frequency analysis was performed on the source reconstructed data with the same analysis as in the sensor space analysis (wavelet decomposition using a complex 6-cycle Morlet wavelet, logarithmically stepped between 0.5 and 15 Hz). For the TRF analyses, the TRF function estimated at the sensor level was used as input into the source reconstruction and no further time-frequency analysis was performed.

Individual participant brains were averaged onto a common participant brain using the FreeSurfer Suite using a transformation based on Montreal Neurological Institute coordinates. Participant data were smoothed across participants (on the participant average brain) using a Gaussian smoothing function.

## Results

### Behavioral results

Task performance was generally good, indicating that the participants were attending appropriately according to instructions. In the Single Speaker conditions, performance was equally good in the A and AV trials (hit rates between 78 and 80%; $t_{(12)} = 1.12$, $p > 0.2$); however, in the Cocktail Party conditions performance was significantly reduced in the A trials compared with the AV trials (hit rates of 74 vs 81%, respectively; $t_{(12)} = 2.44$, $p < 0.05$; Fig. 1*B*).

### MEG results

We used two complementary approaches to quantify neural envelope tracking and its selectivity in the Cocktail Party condition. The first phase dissimilarity/selectivity approach evaluates the consistency of the neural responses across repetitions of the same stimulus, whereas the second temporal response function estimation approach computes the direct relationship between the stimulus and the neural response it elicits.

*Phase dissimilarity and selectivity*

Figure 2*A* illustrates qualitatively that listening to different speech tokens in the Single Speaker condition elicits markedly different temporal patterns of the neural response, as previously demonstrated (Luo and Poeppel, 2007). This effect was quantified by calculating the phase-dissimilarity index, separately for A and AV Single Speaker trials. Significant phase dissimilarity was found for frequencies between 2 and 10 Hz ($p < 0.01$; Fig 3*A*) in both conditions, indicating that the phase of neural activity in this frequency range faithfully tracked the presented speech token. The spatial distribution of the phase-dissimilarity index in

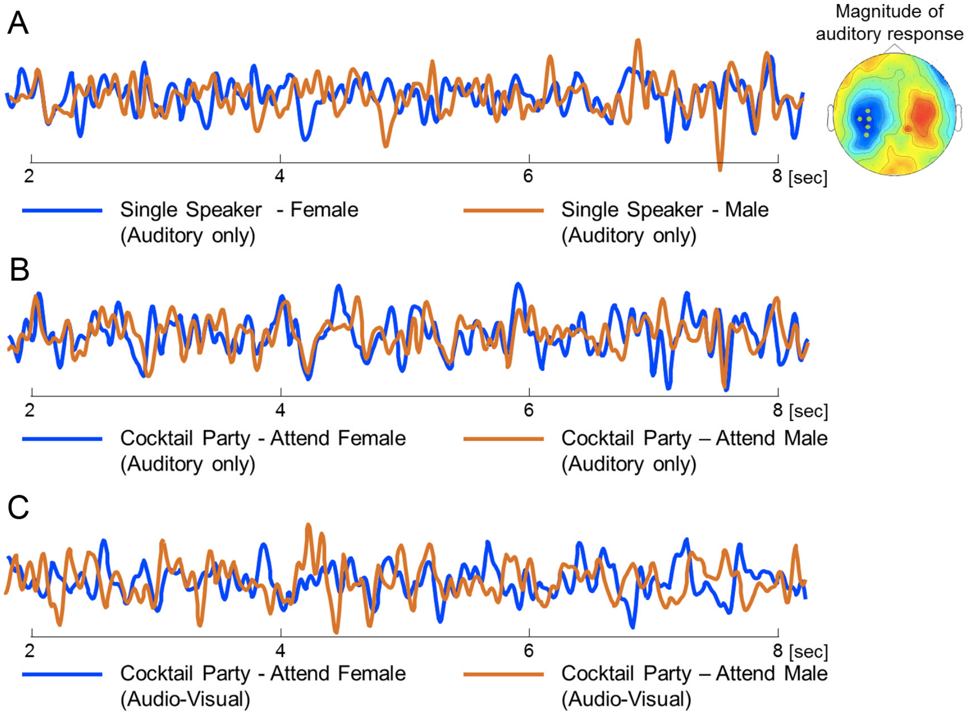

**Figure 2.** Time course of MEG responses. Example time courses of the neural response to the different speech tokens, averaged over trials and participants and filtered between 1 and 10 Hz. Data are shown averaged over the five sensors with the strongest auditory response (with negative polarity), indicated by green dots in the topographical map on the right. ***A***, Neural response to two different tokens presented Auditory Single Speaker condition. Replicating previous findings (Luo and Poeppel, 2007), different speech tokens elicit strikingly different time courses of response. ***B***, Neural responses when attending either to the female or male speaker in the Auditory Cocktail condition. The time courses in the two attentional conditions overlap significantly, with no apparent attentional modulation. ***C***, Neural responses when attending either to the female or male speaker in the Audio-Visual Cocktail condition. In this case, the time courses in the two attentional conditions do not overlap, but rather the temporal pattern of the response is highly influenced by attention, yielding markedly different patterns when attending to different speakers, despite the identical acoustic input.

this frequency range was typical of auditory responses. The difference between the phase dissimilarity in the A and AV conditions was not significant ($t < 1.0$), indicating that adding visual input did not significantly improve the representation of the stimulus in auditory cortex.

In the Cocktail Party conditions we calculated a phase-selectivity index, which follows the same logic as the phase-dissimilarity index but characterizes how different the temporal pattern of the neural responses is when attention is directed toward different speakers (attend female vs attend male), even though the acoustic input remains the same (combination of the two voices). In this case, high phase selectivity indicates selective representation (tracking) of the attended stimulus, whereas low phase selectivity indicates similar patterns of the neural responses, despite attending to different speakers; this pattern likely represents a mixture of responses to both speakers with no detectable modulation by attention. Examples of how allocating attention to different speakers influenced the time course of the neural response in the A and AV cocktail conditions are shown in Figure 2*B*,*C*.

Significant phase selectivity was found in the AV cocktail condition between 3 and 8 Hz (Fig 3*B*) and shared a similar auditory spatial distribution as the phase dissimilarity in the Single Speaker condition. Crucially, the phase selectivity in the A cocktail condition did not reach the threshold for significance at any frequency, and was significantly lower than in the AV cocktail condition ($t_{(12)} = 3.07, p < 0.01$).

The implication of these results is that viewing the face of an attended speaker in a Cocktail Party situation enhances the capacity of auditory cortex to selectively represent and track that speaker, just as if that speaker were presented alone. However, this capacity is sharply reduced when relying only on auditory information alone, with the neural response showing no detectable selectivity.

We next investigated whether the observed envelope tracking in the A and AV conditions indeed reflects activity in auditory cortex, or whether when viewing movies a similar tracking response is found in visual cortex (or both). To this end, we repeated the analysis in source space for five of the participants for whom structural MRIs were available. Results show that the phase-dissimilarity and phase-selectivity effects are entirely attributed to auditory cortex, in both the A and AV conditions (Fig. 4). The source localization results did not adhere to strict anatomical boundaries (e.g., the Sylvian Fissure), but rather formed a region of activity centered in auditory regions but extending beyond these areas. This pattern is likely due to limitations of the source reconstruction given the low number of participants for this analysis (five), as well as field spread of low-frequency signals (see discussion by Schoffelen and Gross, 2009; Peelle et al., 2012).

### *TRF estimation*

The phase-dissimilarity index gives a robust estimate of how well the neural response represents a particular stimulus, but it remains an indirect measure of envelope tracking. A more direct approach is to estimate a TRF, which models the relationship between speech stimuli and the neural responses they elicit, as well as the temporal lag between them (Theunissen et al., 2001; Lalor and Foxe, 2010; Ding and Simon, 2012). We estimated the TRFs between the speech envelope and the neural response in each of the four conditions. In the Cocktail Party conditions, the

TRF model was estimated using the envelopes of both the attended and ignored speakers, which allowed us to compare the responses to each of the concurrently presented stimuli and estimate the relative contribution of each stimulus to the measured response (see Materials and Methods). The predictive power of the model (averaged across all conditions) was significantly higher than chance ($t_{(12)} = 3.56$, $p < 0.005$ vs surrogate TRFs), confirming the robustness of the estimation.

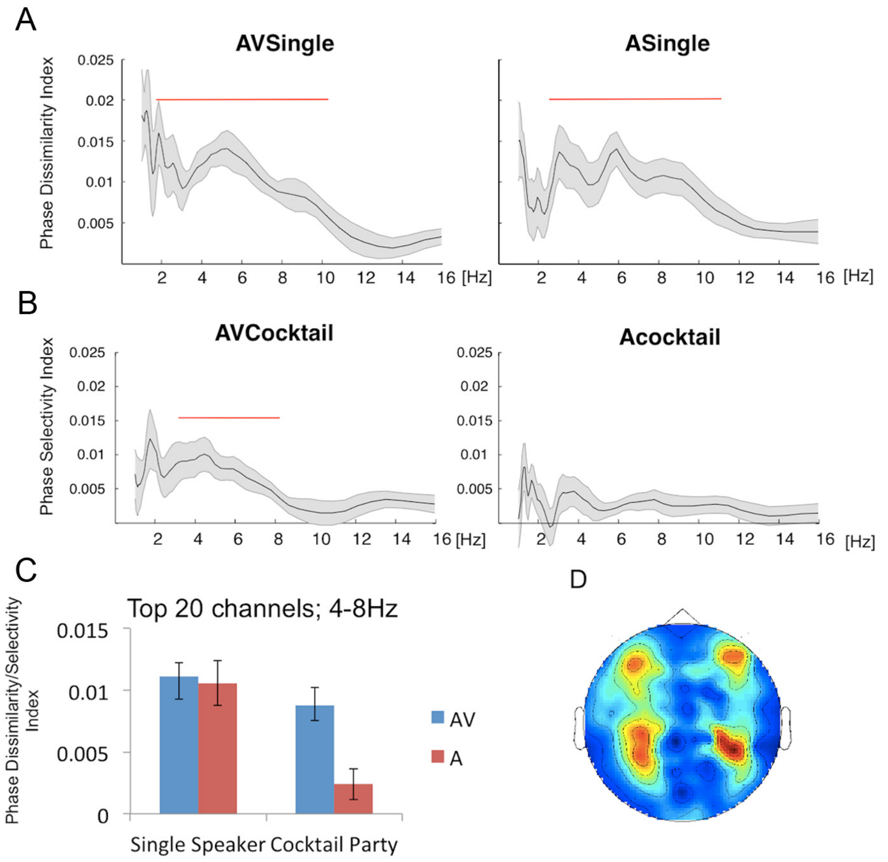

**Figure 3.** Phase dissimilarity and selectivity. ***A***, Phase-dissimilarity values at frequencies between 1 and 16 Hz for the Single Speaker trials in the AV (left) and A (right) conditions, averaged across participants at the top 20 channels (gray shadow indicates ±1 SEM over participants). The red line indicates frequencies where phase dissimilarity was significant in each condition ($p < 0.01$) ***B***, Phase selectivity values between 1 and 16 Hz for the Cocktail Party trials in the AV (left) and A (right) conditions, averaged across participants at the top 20 channels (gray shadow indicates ±1 SEM over participants). The red line indicates frequencies where phase selectivity was significant in each condition ($p < 0.01$) ***C***, Average values of phase dissimilarity/selectivity between 4 and 8 Hz in all conditions. ***D***, Topographical distribution of phase dissimilarity/selectivity values, averaged across all conditions and participants. This distribution is typical of auditory responses.

Results of the TRF analysis are shown in Figure 5. The TRF, averaged across all conditions, had a significant peak at ~50 ms ($p < 0.01$ vs surrogate TRFs), which displayed the spatial distribution of an auditory response, indicating that the neural response in auditory cortex tracked the speech envelope with a lag of ~50 ms. The absolute value of this peak was taken to reflect the strength of the envelope tracking, which we compared across the different conditions. In the Single Speaker condition, TRF amplitude was significantly higher for AV versus A trials ($t_{(12)} = 2.7$, $p < 0.05$). Since the phase dissimilarity was significant for both AV and A trials, indicating that different tokens elicit unique temporal patterns of neural response in both conditions, this additional TRF effect could either reflect visual enhancement of the amplitude neural response or improved temporal tracking of the speaker in the AV condition.

In the Cocktail Party condition, we performed a two-way ANOVA between Modality (A/AV) and Speaker (attended/ignored). There was a main effect of Modality ($F_{(1,12)} = 11.9$, $p < 0.005$), which shows that tracking was overall better for AV versus A stimuli. There was also a main effect of Type ($F_{(1,12)} = 5.6$, $p < 0.05$), and the interaction between the factors trended toward significance ($F_{(1,12)} = 3.4$, $p = 0.08$). *Post hoc* analyses indicated that for the AV stimuli, TRF amplitude was significantly higher for the attended versus the ignored speaker ($t_{(13)} = 2.6$, $p < 0.05$), but this difference was not significant in the A trials ($t < 1.0$). Source reconstruction of the TRF signal in a subset of five participants confirmed that the envelope-tracking response originated in auditory cortex (Fig 5*B*). No evidence for an envelope-tracking response was found in visual regions.

## Discussion

The findings demonstrate that viewing a speaker's face enhances the capacity of auditory cortex to track the temporal speech envelope of that speaker. Visual facilitation is most effective in a Cocktail Party setting, and promotes preferential or selective tracking of the attended speaker, whereas without visual input no significant preference for the attended is achieved. This pattern is in line with behavioral studies showing that the contribution of congruent visual input to speech processing is most substantial under noisy auditory conditions (O'Neill, 1954; Sumby and Pollack, 1954; Helfer and Freyman, 2005; Ross et al., 2007; Bishop and Miller, 2009). These results underscore the importance of visual input in resolving perceptual ambiguity and in directing attention toward a behaviorally relevant speaker in a noisy environment.

Behavioral studies also show robust interference effects from ignored stimuli (Cherry, 1953; Moray, 1959; Wood and Cowan, 1995; Beaman et al., 2007), which are, arguably, due to the fact that ignored stimuli are also represented in auditory cortex (albeit often with reduced amplitude; Woldorff et al., 1993; Ding and Simon, 2012). Thus, improving the selectivity of auditory tracking is likely to have causal implications for performance, and indeed here we show that in the Cocktail Party conditions performance improved in the AV compared with the A condition, alongside an increase in selective tracking.

### Visual cues facilitate speech-envelope tracking

Multiple studies have shown that congruent visual input enhances the neural responses to speech in auditory cortex and in higher order speech-related areas (Callan et al., 2003; Sekiyama et al., 2003; Besle et al., 2004; Bishop and Miller, 2009; McGettigan et al., 2012); however, exactly how the visual input influences activity in auditory cortex is not well understood. In this study we find that when a Single Speaker is presented, reliable low-frequency phase tracking is achieved in

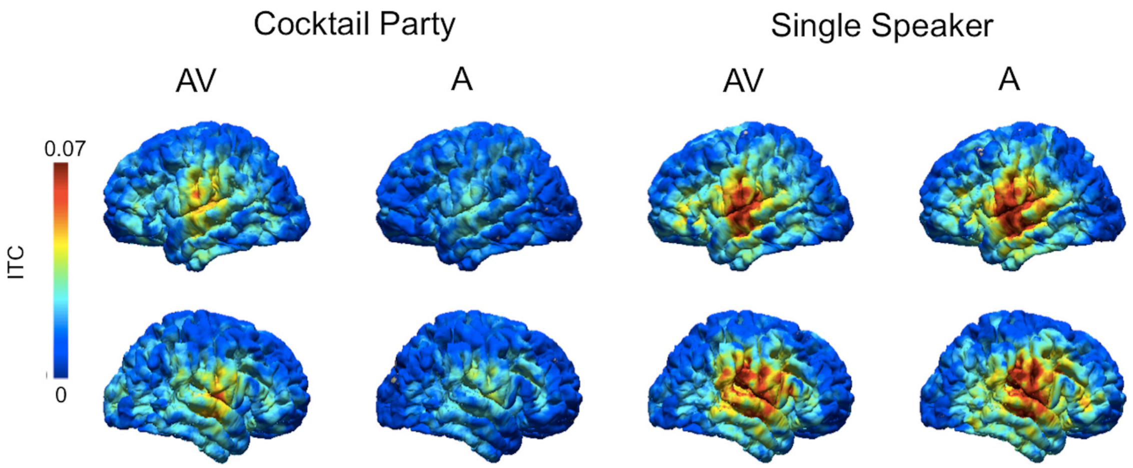

**Figure 4.** Source reconstruction of the phase dissimilarity/selectivity. Reconstruction of ITC between 4 and 8 Hz was performed in each condition for a subset of five participants. Hot and cold colors, respectively, represent strong and weak ITC. Results indicate that phase tracking originates in auditory cortex in all conditions, but is substantially reduced in the A Cocktail condition. Notably, no evidence for phase tracking is found in visual cortex in either of the AV conditions.

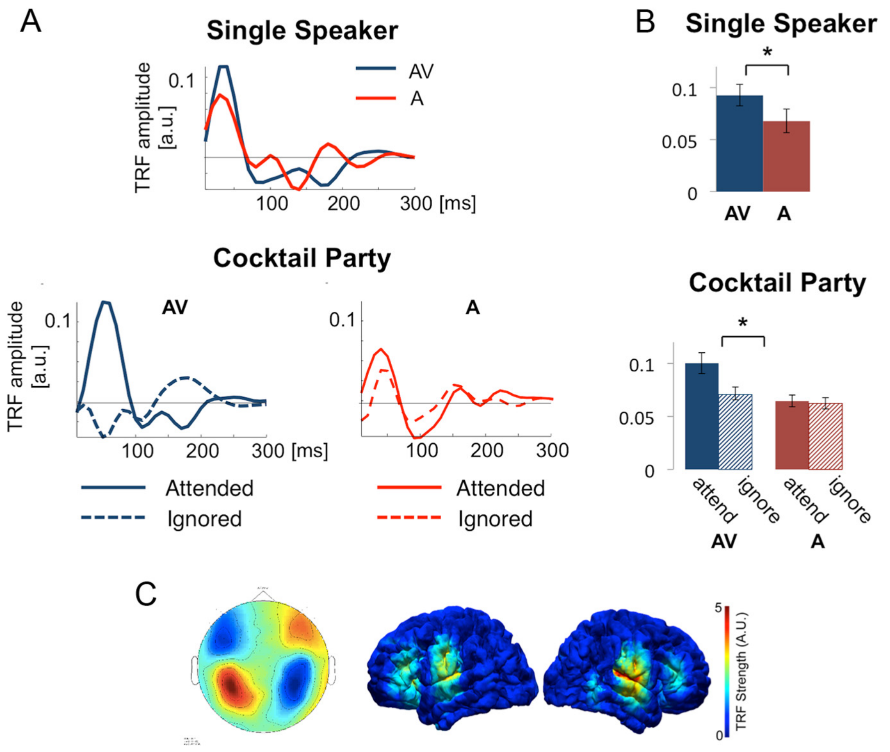

**Figure 5.** TRF analysis. ***A***, Estimated TRF waveforms across all conditions, averaged over the top 10 sensors with positive TRF peak polarity. Top, TRFs to AV and A speech in the Single Speaker condition. The TRFs share a similar time course, with a prominent peak at 50 ms, which is larger in the AV versus A condition. Bottom, TRFs to attended and ignored speakers in the AV (left) and A (right) conditions. In the AV condition the response is strikingly selective for the attended speaker whereas in the A condition similar responses are found for attended and ignored speakers. ***B***, Bar graphs depicting the average TRF peak amplitude at 50 ms across all conditions (absolute value), averaged across the top 20 sensors. ***C***, Topographical distribution of TRF peak amplitude averaged over all participants and conditions (left) and source reconstruction of the TRF peak from five participants (right), indicating it originates in auditory cortex.

auditory cortex both with and without visual input (as indicated by equivalent degrees of phase dissimilarity), yet the amplitude of this tracking response is enhanced in the AV condition (shown by a larger TRF peak amplitude), implying visual amplification of the auditory response. In the Cocktail Party case the role of visual input becomes more crucial since in its absence similar responses are obtained for attended and ignored speakers yielding no significant selectivity, whereas viewing the talking face allows auditory cortex to preferentially track the attended at the expense of the ignored speaker.

Which aspects of the visual input facilitate envelope tracking in auditory cortex? Two types of facial gestures in speech, which operate on different timescales, have been shown to correlate with speech acoustics and improve speech processing. The first are articulation movements of the mouth and jaw, which are correlated with the temporal envelope of speech, both of which are temporally modulated at rates between 2 and 7 Hz (Grant and Seitz, 2000; Chandrasekaran et al., 2009), commensurate with the syllabic rate of speech and the frequency range where we and others have found phase-tracking effects. Several studies have demonstrated increased speech recognition and intelligibility when acoustic input was accompanied by visualization of articulatory gestures (Grant and Seitz, 2000; Grant, 2001; Kim and Davis, 2003). Beyond the contribution of articulatory gestures, other aspects of body movements, such as head and eyebrow movements, have also been shown to improve speech recognition (Munhall et al., 2004; Scarborough et al., 2009). Head movements and other body motions are linked to the production of suprasegmental features of speech such as stress, rhythmicity, and other aspects of prosody (Birdwhistell, 1970; Bennett, 1980; Hadar et al., 1983).

Common to both types of facial gestures is the fact that they precede the acoustic signal. Facial articulation movements precede speech by 100–300 ms (Grant and Seitz, 2000; Chandrasekaran et al., 2009), and the onset of head movements generally precedes the onset of stressed syllables by at least 100 ms (Hadar et al., 1983). This has lead to the suggestion that visual input assists speech perception by predicting the timing of the upcoming auditory input. Estimating when auditory input will occur serves to enhance sensitivity and promote optimal processing in auditory cortex. Supporting this proposal, Schwartz et al. (2004) demonstrated improved intelligibility when auditory syllables were presented together with lip movements that predicted the timing of auditory input, even if the visual cues themselves carried no information about the identity of the syllable (Kim and Davis, 2003). Further supporting the predictive role carried by visual input, neurophysiological data by Arnal et al. (2009, 2011) demonstrate that the early neural response to AV syllables (<200 ms) is enhanced in a manner proportional to the predictive value carried by the visual inpu**t.**

The predictive role assigned to visual cues is in line with the "Attention in Time" hypothesis (Large and Jones, 1999; Jones et al., 2006; Nobre et al., 2007; Nobre and Coull, 2010), which posits that attention can be directed to particular points in time when relevant stimuli are expected, similar to the allocation of spatial attention. Speech is naturally rhythmic, and predictions about the timing of upcoming events (e.g., syllables) can be formed from temporal regularities within the acoustics alone (Elhilali et al., 2009; Shamma et al., 2011), yet visual cues that precede the audio can serve to reinforce and tune those predictions. Moreover, if the visual input also carries predictive informative about the speech content (say a bilabial vs velar articulation), the listener can derive further processing benefit.

### A mechanistic perspective on speech-envelope tracking

From a mechanistic perspective, we can offer two hypotheses as to how the Attention in Time hypothesis could be implemented on the neural level. The first possibility is that auditory cortex contains spectrotemporal representations of both speakers; however, the portion of the auditory response that is temporally coherent with the visual input is selectively amplified. This perspective of binding through temporal coherence is in line with computational perspectives on stream segregation (Elhilali et al., 2009). Alternatively, it has been shown that predictive nonauditory input can reset the phase of low-frequency oscillations in auditory cortex (Lakatos et al., 2007, 2009; Kayser et al., 2008), a mechanism that could be particularly advantageous for improving selective envelope tracking under adverse auditory conditions, such as the Cocktail Party situation. Since low-frequency oscillations govern the timing of neuronal excitability (Buzsáki and Chrobak, 1995; Lakatos et al., 2005; Mizuseki et al., 2009), visually guided phase resets in auditory cortex would align the timing of high neuronal excitability with the timing of attended events, and consequently, events from the to-be-ignored streams would naturally fall on random phases of excitability, contributing to their perceptual attenuation (Schroeder et al., 2008; Zion Golumbic et al., 2012). Both perspectives emphasize the significance of temporal structure in forming a neural representation for speech and selecting the appropriate portion of the auditory scene; however, additional research is needed to fully understand the mechanistic interaction between the visual and auditory speech content.

Another mechanistic question raised by these data are whether visual facilitation of the envelope-tracking response is mediated through multisensory regions, such as the superior temporal and intraparietal sulci, and then fed back to auditory cortex (Beauchamp et al., 2004), or whether it is brought about through feedforward projections from extralemniscal thalamic regions, or direct lateral connections between visual and auditory cortex (Schroeder et al., 2008; Musacchia and Schroeder, 2009). Adjudicating between these possibilities requires additional research; however, the fact that AV influences were observed here as early as 50 ms hints that it is influenced through either thalamocortical or lateral connections between the sensory cortices.

### Relationship to previous studies

There is evidence that while watching movies, visual cortex also displays phase locking to the stimulus (Luo et al., 2010). However, in the current study both the phase-dissimilarity/selectivity effects and TRF estimation were localized to auditory cortex, and we found no evidence for phase locking and/or envelope tracking in visual regions. This does not preclude the possibility that there is phase locking to the videos in visual cortex, which might be too weak to pick up using the current experimental design (due to the relatively dull visual input of a face with little motion) or might not be locked to the acoustic envelope.

Previous studies have also shown preferential tracking of attended versus ignored speech even without visual input (Kerlin et al., 2010; Ding and Simon, 2012), similar to classic attentional modulation of evoked responses (Hubel et al., 1959; Hillyard et al., 1973; Tiitinen et al., 1993). These findings contrast with the current results where we failed to find significant attentional modulations in the auditory-only condition. However, critically, in those studies the two speakers were presented from different spatial locations, and not from a central location as in the current study. It is well established that spatial information contributes to stream segregation and attentional selection (Moore and Gockel, 2002; Fritz et al., 2007; Hafter et al., 2007; Elhilali and Shamma, 2008; Elhilali et al., 2009; McDermott, 2009; Shamma et al., 2011). Thus, the attentional effects reported in those studies could have been influenced by spatial cues. Future experiments are needed to assess the relative contribution of visual, spatial, and spectral cues to selective envelope-tracking.

## References

Aiken SJ, Picton TW (2008) Human cortical responses to the speech envelope. Ear Hear 29:139–157. 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

Arnal LH, Wyart V, Giraud AL (2011) Transitions in neural oscillations reflect prediction errors generated in audiovisual speech. Nat Neurosci 14:797–801. CrossRef Medline

Beaman CP, Bridges AM, Scott SK (2007) From dichotic listening to the irrelevant sound effect: a behavioural and neuroimaging analysis of the processing of unattended speech. Cortex 43:124–134. 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

Bennett AT (1980) Rhythmic analysis of multiple levels of communicative behavior in face to face interaction. In: Aspects of nonverbal communication (Raffler-Engel W, ed), pp 237–251. Lisse, the Netherlands: Swets and Zeitleinger.

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

Birdwhistell RL (1970) Kinesics and context: essays on body motion communication. Philadelphia: University of Pennsylvania.

Bishop CW, Miller LM (2009) A multisensory cortical network for understanding speech in noise. J Cogn Neurosci 21:1790–1805. Medline

Buzsáki G, Chrobak JJ (1995) Temporal structure in spatially organized neuronal ensembles: a role for interneuronal networks. Curr Opin Neurobiol 5:504–510. CrossRef Medline

Callan D, Jones JA, Munhall K, Callan AM, Kroos C, Vatikiotis-Bateson E (2003) Neural processes underlying perceptual enhancement by visual speech gestures. Neuroreport 14:2213–2218. 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

Cherry EC (1953) Some experiments on the recognition of speech, with one and two ears. J Acoust Soc Am 25:975–979. CrossRef

David SV, Mesgarani N, Shamma SA (2007) Estimating sparse spectro-temporal receptive fields with natural stimuli. Network 18:191–212. CrossRef Medline

Davis C, Kislyuk D, Kim J, Sams M (2008) The effect of viewing speech on auditory speech processing is different in the left and right hemispheres. Brain Res 1242:151–161. CrossRef Medline

de Cheveigné A, Simon JZ (2007) Denoising based on time-shift pca. J Neurosci Methods 165:297–305. 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

Drullman R (2006) The significance of temporal modulation frequencies for speech intelligibility. In: Listening to speech: an auditory perspective. (Greenberg S, Ainsworth W, eds), pp 39–48. Mahwah, NJ: Lawrence Erlbaum.

Elhilali M, Shamma SA (2008) A cocktail party with a cortical twist: how cortical mechanisms contribute to sound segregation. J Acoust Soc Am 124:3751–3771. 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

Fritz JB, Elhilali M, David SV, Shamma SA (2007) Auditory attention–focusing the searchlight on sound. Curr Opin Neurobiol 17:437–455. CrossRef Medline

Ghitza O (2011) Linking speech perception and neurophysiology: speech decoding guided by cascaded oscillators locked to the input rhythm. Front Psychol 2:130. 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. 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

Hadar U, Steiner TJ, Grant EC, Rose FC (1983) Head movement correlates of juncture and stress at sentence level. Lang Speech 26:117–129. Medline

Hafter E, Sarampalis A, Loui P (2007) Auditory attention and filters. In: Auditory perception of sound sources. (Yost W, Popper AN, Fay RR, eds), pp 115–142. New York: Springer.

Hämäläinen MS, Ilmoniemi RJ (1994) Interpreting magnetic fields of the brain: minimum norm estimates. Med Biol Eng Comput 32:35–42. CrossRef Medline

Helfer KS, Freyman RL (2005) The role of visual speech cues in reducing energetic and informational masking. J Acoust Soc Am 117:842–849. CrossRef Medline

Hertrich I, Dietrich S, Trouvain J, Moos A, Ackermann H (2012) Magnetic brain activity phase-locked to the envelope, the syllable onsets, and the fundamental frequency of a perceived speech signal. Psychophysiology 49:322–334. CrossRef Medline

Hillyard SA, Hink RF, Schwent VL, Picton TW (1973) Electrical signs of selective attention in the human brain. Science 182:177–180. CrossRef Medline

Hubel DH, Henson CO, Rupert A, Galambos R (1959) "Attention" units in the auditory cortex. Science 129:1279–1280. CrossRef Medline

Jones MR, Johnston HM, Puente J (2006) Effects of auditory pattern structure on anticipatory and reactive attending. Cogn Psychol 53:59–96. CrossRef Medline

Jung TP, Makeig S, Westerfield M, Townsend J, Courchesne E, Sejnowski TJ (2000) Removal of eye activity artifacts from visual event-related potentials in normal and clinical subjects. Clin Neurophysiol 111: 1745–1758. CrossRef Medline

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

Kerlin JR, Shahin AJ, Miller LM (2010) Attentional gain control of ongoing cortical speech representations in a "cocktail party." J Neurosci 30:620–628. CrossRef Medline

Kim J, Davis C (2003) Hearing foreign voices: does knowing what is said affect visual-masked-speech detection? Perception 32:111–120. CrossRef Medline

Lakatos P, Shah AS, Knuth KH, Ulbert I, Karmos G, Schroeder CE (2005) An oscillatory hierarchy controlling neuronal excitability and stimulus processing in the auditory cortex. J Neurophysiol 94:1904–1911. 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

Lakatos P, O'Connell MN, Barczak A, Mills A, Javitt DC, Schroeder CE (2009) The leading sense: supramodal control of neurophysiological context by attention. Neuron 64:419–430. CrossRef Medline

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

Large EW, Jones MR (1999) The dynamics of attending: how people track time-varying events. Psychological Rev 106:119–159. CrossRef

Luo H, Poeppel D (2007) Phase patterns of neuronal responses reliably discriminate speech in human auditory cortex. Neuron 54:1001–1010. 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

McDermott JH (2009) The cocktail party problem. Curr Biol 19: R1024–R1027. CrossRef Medline

McGettigan C, Faulkner A, Altarelli I, Obleser J, Baverstock H, Scott SK (2012) Speech comprehension aided by multiple modalities: behavioural and neural interactions. Neuropsychologia 50:762–776. CrossRef Medline

Mesgarani N, Chang EF (2012) Selective cortical representation of attended speaker in multi-talker speech perception. Nature 485:233–236. CrossRef Medline

Mizuseki K, Sirota A, Pastalkova E, Buzsáki G (2009) Theta oscillations provide temporal windows for local circuit computation in the entorhinal-hippocampal loop. Neuron 64:267–280. CrossRef Medline

Moore BCJ, Gockel H (2002) Factors influencing sequential stream segregation. Acta Acustica 88:320–332.

Moray N (1959) Attention in dichotic listening: affective cues and the influence of instructions. Q J Exp Psychol 11:56–60. CrossRef

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

Musacchia G, Schroeder CE (2009) Neuronal mechanisms, response dynamics and perceptual functions of multisensory interactions in auditory cortex. Hear Res 258:72–79. CrossRef Medline

Nobre AC, Coull JT (2010) Attention and time. Oxford, UK: Oxford UP.

Nobre A, Correa A, Coull J (2007) The hazards of time. Curr Opin Neurobiol 17:465–470. CrossRef Medline

Oldfield RC (1971) The assessment and analysis of handedness: the Edinburgh inventory. Neuropsychologia 9:97–113. CrossRef Medline

O'Neill JJ (1954) Contributions of the visual components of oral symbols to speech comprehension. J Speech Hearing Disord 19:429–439. Medline

Oostenveld R, Fries P, Maris E, Schoffelen J-M (2011) Fieldtrip: open source software for advanced analysis of MEG, EEG, and invasive electrophysiological data. Comput Intel Neurosci 2011.

Peelle JE, Gross J, Davis MH (2012) Phase-locked responses to speech in human auditory cortex are enhanced during comprehension. Cereb Cortex. Advance online publication. Retrieved May 17, 2012. doi:10.1093/cercor/bhs118. CrossRef

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 (2007) Do you see what i am saying? Exploring visual enhancement of speech comprehension in noisy environments. Cereb Cortex 17:1147–1153. Medline

Scarborough R, Keating P, Mattys SL, Cho T, Alwan A (2009) Optical phonetics and visual perception of lexical and phrasal stress in English. Lang Speech 52:135–175. CrossRef Medline

Schoffelen JM, Gross J (2009) Source connectivity analysis with MEG and EEG. Hum Brain Mapp 30:1857–1865. 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, Berthommier F, Savariaux C (2004) Seeing to hear better: evidence for early audio-visual interactions in speech identification. Cognition 93:B69–B78. CrossRef Medline

Sekiyama K, Kanno I, Miura S, Sugita Y (2003) Auditory-visual speech perception examined by fMRI and PET. Neurosci Res 47:277–287. CrossRef Medline

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

Shannon RV, Zeng FG, Kamath V, Wygonski J, Ekelid M (1995) Speech recognition with primarily temporal cues. Science 270:303–304. CrossRef Medline

Studebaker GA (1985) A "rationalized" arcsine transform. J Speech Hear Res 28:455–462. Medline

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

Theunissen FE, David SV, Singh NC, Hsu A, Vinje WE, Gallant JL (2001) Estimating spatio-temporal receptive fields of auditory and visual neurons from their responses to natural stimuli. Network 12:289–316. CrossRef Medline

Tiitinen H, Sinkkonen J, Reinikainen K, Alho K, Lavikainen J, Näätänen R (1993) Selective attention enhances the auditory 40-hz transient response in humans. Nature 364:59–60. 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

Woldorff MG, Gallen CC, Hampson SA, Hillyard SA, Pantev C, Sobel D, Bloom FE (1993) Modulation of early sensory processing in human auditory cortex during auditory selective attention. Proc Natl Acad Sci U S A 90:8722–8726. CrossRef Medline

Wood N, Cowan N (1995) The cocktail party phenomenon revisited: how frequent are attention shifts to one's name in an irrelevant auditory channel? J Exp Psychol Learn Mem Cogn 21:255–260. CrossRef Medline

Zion Golumbic EM, Poeppel D, Schroeder CE (2012) Temporal context in speech processing and attentional stream selection: a behavioral and neural perspective. Brain Lang 122:151–161. CrossRef Medline