

Cortical Representations of Speech in a Multitalker Auditory Scene

Krishna C. Puvvada¹ and  Jonathan Z. Simon^{1,2,3}

¹Department of Electrical & Computer Engineering, ²Department of Biology, and ³Institute for Systems Research, University of Maryland, College Park, Maryland 20742

The ability to parse a complex auditory scene into perceptual objects is facilitated by a hierarchical auditory system. Successive stages in the hierarchy transform an auditory scene of multiple overlapping sources, from peripheral tonotopically based representations in the auditory nerve, into perceptually distinct auditory-object-based representations in the auditory cortex. Here, using magnetoencephalography recordings from men and women, we investigate how a complex acoustic scene consisting of multiple speech sources is represented in distinct hierarchical stages of the auditory cortex. Using systems-theoretic methods of stimulus reconstruction, we show that the primary-like areas in the auditory cortex contain dominantly spectrotemporal-based representations of the entire auditory scene. Here, both attended and ignored speech streams are represented with almost equal fidelity, and a global representation of the full auditory scene with all its streams is a better candidate neural representation than that of individual streams being represented separately. We also show that higher-order auditory cortical areas, by contrast, represent the attended stream separately and with significantly higher fidelity than unattended streams. Furthermore, the unattended background streams are more faithfully represented as a single unsegregated background object rather than as separated objects. Together, these findings demonstrate the progression of the representations and processing of a complex acoustic scene up through the hierarchy of the human auditory cortex.

Key words: attention; auditory cortex; cocktail party problem; magnetoencephalography; stimulus reconstruction; temporal response function

Significance Statement

Using magnetoencephalography recordings from human listeners in a simulated cocktail party environment, we investigate how a complex acoustic scene consisting of multiple speech sources is represented in separate hierarchical stages of the auditory cortex. We show that the primary-like areas in the auditory cortex use a dominantly spectrotemporal-based representation of the entire auditory scene, with both attended and unattended speech streams represented with almost equal fidelity. We also show that higher-order auditory cortical areas, by contrast, represent an attended speech stream separately from, and with significantly higher fidelity than, unattended speech streams. Furthermore, the unattended background streams are represented as a single undivided background object rather than as distinct background objects.

Introduction

Individual sounds originating from multiple sources in a complex auditory scene mix linearly and irreversibly before they enter the ear, yet are perceived as distinct objects by the listener

(Cherry, 1953; Bregman, 1994; McDermott, 2009). The separation, or rather individual re-creation, of such linearly mixed original sound sources is a mathematically ill-posed problem, yet the brain nevertheless routinely performs this task with ease. The neural mechanisms by which this perceptual “unmixing” of sounds occur, the collective cortical representations of the auditory scene and its constituents, and the role of attention in both are key problems in contemporary auditory neuroscience.

It is known that auditory processing in the primate cortex is hierarchical (Davis and Johnsrude, 2003; Hickok and Poeppel, 2007; Rauschecker and Scott, 2009; Okada et al., 2010; Peelle et al., 2010; Overath et al., 2015) with subcortical areas projecting onto the core areas of the auditory cortex, and from there on to belt, parabelt, and additional auditory areas (Kaas and Hackett,

Received April 6, 2017; revised July 20, 2017; accepted Aug. 8, 2017.

Author contributions: K.C.P. and J.Z.S. designed research; K.C.P. and J.Z.S. performed research; K.C.P. and J.Z.S. analyzed data; K.C.P. and J.Z.S. wrote the paper.

This work was supported by National Institutes of Health Grant R01 DC014085. We thank Elizabeth Nguyen for excellent MEG technical support, and Stuart Rosen for advice on the stimulus paradigm.

The authors declare no competing financial interests.

Correspondence should be addressed to Jonathan Z. Simon, Department of Electrical & Computer Engineering, University of Maryland, 8223 Paint Branch Drive, College Park, MD 20742. E-mail: jzsimon@umd.edu.

DOI:10.1523/JNEUROSCI.0938-17.2017

Copyright © 2017 the authors 0270-6474/17/379189-08\$15.00/0

2000). Sound entering the ear reaches different anatomical/functional areas of the auditory cortex with different latencies (Recanzone et al., 2000; Nourski et al., 2014). Due to this serial component of auditory processing, the hierarchy of processing can be described by both anatomy and latency, of which the latter may be exploited using the high temporal fidelity of noninvasive magnetoencephalography (MEG) neural recordings.

In selective listening experiments using natural speech and MEG, the two major neural responses known to track the speech envelope are the M50_{TRF} and M100_{TRF}, with respective latencies of 30–80 and 80–150 ms, of which the dominant neural sources are, respectively, Heschl's gyrus (HG) and the planum temporale (PT; Steinschneider et al., 2011; Ding and Simon, 2012b). The posteromedial HG is the site of the core auditory cortex; the PT contains both belt and parabelt auditory areas (here collectively referred to as higher-order areas; Griffiths and Warren, 2002; Sweet et al., 2005). Hence the earlier neural responses are dominated by the core auditory cortex, and the later are dominated by higher-order areas. To better understand the neural mechanisms of auditory scene analysis, it is essential to understand how the cortical representations of a complex auditory scene change from the core to the higher-order auditory areas.

One topic of interest is whether the brain maintains distinct neural representations for each unattended source (in addition to the representation of the attended source), or whether all unattended sources are represented collectively as a single monolithic background object. A common paradigm used to investigate the neural mechanisms underlying auditory scene analysis employs a pair of speech streams, of which one is attended, which then leaves the other speech stream remaining as the background (Kerlin et al., 2010; Ding and Simon, 2012b; Mesgarani and Chang, 2012; Power et al., 2012; Zion Golumbic et al., 2013b; O'Sullivan et al., 2015). This results in a limitation that cannot address the question of distinct versus collective neural representations for unattended sources. This touches on the long-standing debate of whether auditory object segregation is preattentive or is actively influenced by attention (Carlyon, 2004; Sussman et al., 2005; Shinn-Cunningham, 2008; Shamma et al., 2011). Evidence for segregated neural representations of background streams would support the former, whereas a lack of segregated background objects would support the latter.

To address these issues, we use MEG to investigate a variety of potential cortical representations of the elements of a multitalker auditory scene. We test two major hypotheses: that the dominant representation in the core auditory cortex is of the physical acoustics, not of separated auditory objects; and that once object-based representations emerge in higher-order auditory areas, the unattended contributions to the auditory scene are represented collectively as a single background object. The methodological approach employs the linear systems methods of stimulus prediction and MEG response reconstruction (Lalor et al., 2009; Mesgarani et al., 2009; Ding and Simon, 2012b; Mesgarani and Chang, 2012; Pasley et al., 2012; Di Liberto et al., 2015).

Materials and Methods

Subjects and experimental design. Nine normal-hearing, young adults (six female) participated in the experiment. All subjects were paid for their participation. The experimental procedures were approved by the University of Maryland Institutional Review Board. Subjects listened to a mixture of three speech segments spoken by, respectively, a male adult, a female adult, and a child. The three speech segments were mixed into a single audio channel with equal perceptual loudness. All three speech segments were taken from public domain narration of *Grimms' Fairy*

Tales by Jacob and Wilhelm Grimm (<https://librivox.org/fairy-tales-by-the-brothers-grimm/>). Periods of silence >300 ms were replaced by a shorter gap whose duration was chosen randomly between 200 and 300 ms. The audio signal was low-pass filtered with cutoff at 4 kHz. In the first of three conditions, Condition 1, the subjects were asked to attend to the child speaker, while ignoring the other two (i.e., child speaker as target, with male and female adult speakers as background). In Condition 2, during which the same mixture was played as in Condition 1, the subjects were instead asked to attend to the male adult speaker (with the female adult and child speakers as background). Similarly, in Condition 3, the target was switched to the female adult speaker. Each condition was repeated three times in succession, producing three trials per condition. The presentation order of the three conditions was counterbalanced across subjects. Each trial lasted 220 s, divided into two 110 s partitions, to reduce listener fatigue. To help participants attend to the correct speaker, the first 30 s of each partition was replaced by the clean recording of the target speaker alone, followed by a 5 s upward linear ramp of the background speakers. Recordings of this first 35 s of each partition were not included in any analysis. To further encourage the subjects to attend to the correct speaker, a target word was set before each trial and the subjects were asked to count the number of occurrences of the target word in the speech of the attended speaker. Additionally, after each condition, the subject was asked to summarize the attended narrative. The subjects were required to close their eyes while listening. Before the main experiment, 100 repetitions of a 500 Hz tone pip were presented to each subject to elicit the M100 response, a reliable auditory response occurring ~100 ms after the onset of a tone pip. These data were used to check whether any potential subjects gave abnormal auditory responses, but no subjects were excluded based on this criterion.

Data recording and preprocessing. MEG recordings were conducted using a 160-channel whole-head system (Kanazawa Institute of Technology, Kanazawa, Japan). Its detection coils are arranged in a uniform array on a helmet-shaped surface of the bottom of the dewar, with ~25 mm between the centers of two adjacent 15.5-mm-diameter coils. Sensors are configured as first-order axial gradiometers with a baseline of 50 mm; their field sensitivities are 5 fT/√Hz or better in the white-noise region. Subjects lay horizontally in a dimly lit magnetically shielded room (Yokogawa Electric). Responses were recorded with a sampling rate of 1 kHz with an on-line 200 Hz low-pass filter and 60 Hz notch filter. Three reference magnetic sensors and three vibrational sensors were used to measure the environmental magnetic field and vibrations. The reference sensor recordings were used to reduce environmental noise from the MEG recordings using the time-shift principal component analysis (TS-PCA) method (de Cheveigné and Simon, 2007). Additionally, MEG recordings were decomposed into virtual sensors/components using denoising source separation (DSS; Särelä and Valpola, 2005; de Cheveigné and Simon, 2008; de Cheveigné and Parra, 2014), a blind source separation method that enhances neural activity consistent over trials. Specifically, DSS decomposes the multichannel MEG recording into temporally uncorrelated components, where each component is determined by maximizing its trial-to-trial reliability, measured by the correlation between the responses to the same stimulus in different trials. To reduce the computational complexity, the 157 MEG sensors were reduced, using DSS, to four components in each hemisphere for all further analysis. Also, both stimulus-envelope and MEG responses were band-pass filtered between 1 and 8 Hz (δ and θ bands), which correspond to the slow temporal modulations in speech (Ding and Simon, 2012a, b).

Neural model terminology and notation. As specified in the stimulus description, the subject attends to one of the three speech streams in each condition. Neural models of speech-stream processing can be compared by contrasting the predicted envelope reconstructions of the different models. The envelope of the attended speech stream is referred to as the “foreground” and the envelope of each of the two unattended speech streams is referred to as the “individual background.” In contrast, the envelope of the entire unattended part of the stimulus, comprising both unattended speech streams, is referred to as the “combined background.” The envelope of the entire acoustic stimulus or auditory scene, comprising all three speech streams is referred to as the “acoustic scene.” Thus, if S_a, S_b, S_c are three speech stimuli, $Env(S_a + S_b + S_c)$ is the acoustic scene.

In contrast, the sum of envelopes of three speech streams, $Env(S_a) + Env(S_b) + Env(S_c)$, is referred to as the “sum of streams.” The acoustic scene and the sum of streams are not mathematically equal: even though both are functions of the same stimuli, they differ due to the nonlinear nature of a signal envelope (the linear correlation between the acoustic scene and the sum of streams is typically ~ 0.75). Combination (unsegregated) envelopes, whether of the entire acoustic scene or the background only, can be used to test neural models that do not perform stream segregation. Sums of individual stream envelopes, whether of all streams or just the background streams, can be used to test neural models that process the (segregated) streams in parallel, given that neurally generated magnetic fields add in linear superposition.

Neural responses with latencies $< \sim 85$ ms (typically originating from core auditory areas) are referred to here as “early neural responses” and responses with latencies $> \sim 85$ ms (typically from higher-order auditory areas; Ahveninen et al., 2011; Okamoto et al., 2011; Steinschneider et al., 2011) are referred to as “late neural responses.”

The next two sections describe models of the neural encoding of stimuli into responses, followed by models of the decoding of stimuli from neural responses. Encoding models are presented here first because of their ease of description over decoding models, but in Results the decoding analysis is presented first, since it is the decoding results that inform the new model of encoding.

Temporal response function. In an auditory scene with a single talker, the relation between MEG neural response and the presented speech stimuli can be modeled using a linear temporal response function (TRF) as follows (Eq. 1):

$$r(t) = \sum_{\tau} s(t - \tau) TRF(\tau) + \epsilon(t)$$

where $t = 0, 1, \dots, T$ is time, $r(t)$ is the response from any individual sensor or DSS component, $s(t)$ is the stimulus envelope in decibels, $TRF(t)$ is the TRF itself, and $\epsilon(t)$ is residual response waveform not explained by the TRF model (Ding and Simon, 2012a). The envelope is extracted by averaging the auditory spectrogram (Chi et al., 2005) along the spectral dimension. The TRF is estimated using boosting with 10-fold cross-validation (David et al., 2007). In case of single speech stimuli, the TRF is typically characterized by a positive peak between 30 and 80 ms and a negative peak between 90 and 130 ms, referred to as $M50_{TRF}$ and $M100_{TRF}$ respectively (Ding and Simon, 2012b). Positivity/negativity of the magnetic field is by convention defined to agree with the corresponding electroencephalography (EEG) peaks. Success/accuracy of the linear model is evaluated by how well it predicts neural responses, as measured by the proportion of the variance explained: the square of the Pearson correlation coefficient between the MEG measurement and the TRF model prediction.

In the case of more than one speaker, the MEG neural response, $r(t)$, can be modeled as the sum of the responses to the individual acoustic sources (Ding and Simon, 2012b; Zion Golumbic et al., 2013b), referred to here as the “summation model.” For example, with three speech streams, the neural response would be modeled as follows (Eq. 2):

$$r(t) = \sum_{\tau=0}^{\tau=\tau_2} S_a(t - \tau) TRF_a(\tau) + \sum_{\tau=0}^{\tau=\tau_2} S_b(t - \tau) TRF_b(\tau) + \sum_{\tau=0}^{\tau=\tau_2} S_c(t - \tau) TRF_c(\tau) + \epsilon(t)$$

where $S_a(t)$, $S_b(t)$, and $S_c(t)$ are the envelopes of the three speech streams, and $TRF_a(t)$, $TRF_b(t)$, and $TRF_c(t)$ are the TRFs corresponding to each stream. τ_2 represents the length of TRF. All TRFs in the summation model are estimated simultaneously.

We propose, in addition to the existing summation model, a new encoding model referred to as the “early–late model,” which enables one to incorporate the hypothesis that the early neural responses typically represent the entire acoustic scene, but that the later neural responses differentially represent the separated foreground and background (Eq. 3):

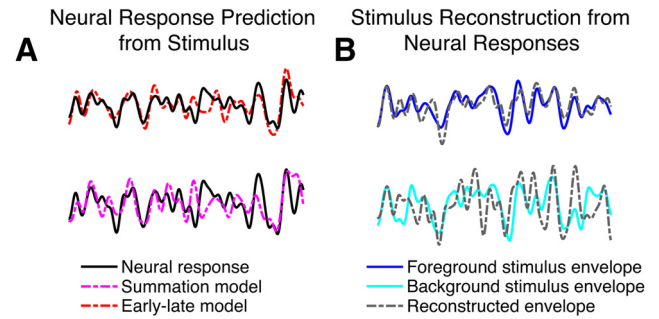


Figure 1. Illustrations of outcomes comparing competing encoding-based and decoding-based neural representations of the auditory scene and its constituents. All examples are grand averages across subjects (3 s duration). **A**, Comparing competing models of encoding to neural responses. In both the top and bottom examples, an experimentally measured MEG response (black) is compared with the neural response predictions made by competing proposed models. In the top example, the neural response prediction (red) is from the early–late model; in the bottom example, the neural response prediction (magenta) is from the summation model. The proposed early–late model prediction shows higher correlation with the actual MEG neural response than the summation model. **B**, Comparing models of decoding to stimulus speech envelopes. In both the top and bottom examples, an acoustic speech stimulus envelope (blue/cyan) is compared with the model reconstruction of the respective envelope (gray). In the top example, the envelope reconstruction is of the foreground stimulus (blue), based on late time responses; in the bottom example, the envelope reconstruction is of the background stimulus (cyan), also based on late time responses. The foreground reconstruction shows higher correlation with the actual foreground envelope, compared with the background reconstruction with the actual background envelope.

$$r(t) = \sum_{\tau=0}^{\tau=\tau_1} S_A(t - \tau) TRF_A(\tau) + \sum_{\tau=\tau_1}^{\tau=\tau_2} S_F(t - \tau) TRF_F(\tau) + \sum_{\tau=\tau_1}^{\tau=\tau_2} S_B(t - \tau) TRF_B(\tau) + \epsilon(t)$$

where $S_A(t)$ is the (entire) acoustic scene, $S_F(t)$ is the envelope of attended (foreground) speech stream, and $S_B(t)$ is the combined background (i.e., envelope of everything other than attended speech stream in the auditory scene), and $TRF_A(t)$, $TRF_F(t)$, and $TRF_B(t)$ are the corresponding TRFs. τ_1 , τ_2 represent the boundary values of the integration windows for early and late neural responses respectively, with $0 < \tau_1 < \tau_2$.

The explanatory power of different models, such as the summation and early–late models, can be ranked by comparing the accuracy of their response predictions (Fig. 1A). The models differ in terms of number of free parameters, with the early–late model having fewer parameters than the summation model. Hence, any improved performance observed in the proposed early–late model over the summation model is correspondingly more likely due to a better model fit, since it has less freedom to fit the data (though the converse would not hold).

Decoding speech from neural responses. While the TRF/encoding analysis described in the previous section predicts neural response from the stimulus, decoding analysis reconstructs the stimulus based on the neural response. Thus, decoding analysis complements the TRF analysis (Mesgarani et al., 2009). Mathematically, the envelope reconstruction/decoding operation can be formulated as follows (Eq. 4):

$$E(t) = \sum_{k=1}^N \sum_{\tau=\tau_b}^{\tau=\tau_e} M_k(t + \tau) D_k(\tau) + \epsilon(t)$$

where $E(t)$ is the stimulus envelope, $M_k(t)$ is the MEG recording (neural response) from sensor/component k , and $D_k(t)$ is the linear decoder for sensor/component k . The times τ_b and τ_e denote the beginning and end times of the integration window. By appropriately choosing the values of τ_b and τ_e , envelope reconstructions using neural responses from any desired time window can be compared. The decoder is estimated using boosting analogously to the TRF estimation in the previous section. In the single-talker case, the envelope is of that talker’s speech. In a multi-

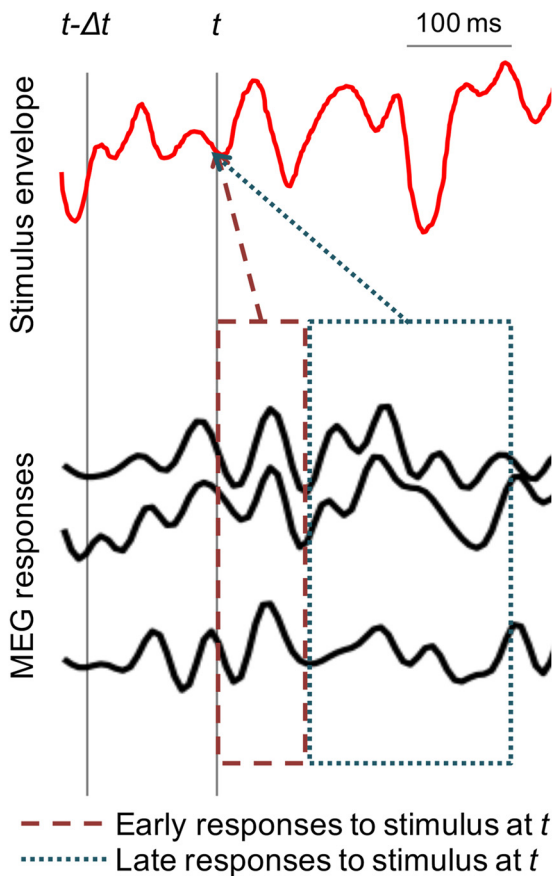


Figure 2. Early versus late MEG neural responses to a continuous speech stimulus. A sample stimulus envelope and time-locked multichannel MEG recordings are shown in red and black respectively. The two gray vertical lines indicate two arbitrary time points at $t - \Delta t$ and t . The dashed and dotted boxes represent the early and late MEG neural responses to stimulus at time point t respectively. The reconstruction of the stimulus envelope at time t can be based on either early or late neural responses, and the separate reconstructions can be compared against each other.

talker case, the envelope to be reconstructed might be the envelope of the speech of the attended talker, or one of the background talkers, or of a combination of any two or all three talkers, depending on the model under consideration. Chance-level reconstruction (i.e., the noise floor) from a particular neural response is estimated by reconstructing an unrelated stimulus envelope from that neural response. Figure 2 illustrates the distinction between reconstruction of the stimulus envelope from early and late responses. The stimulus envelope at time point t can be reconstructed using neural responses from the dashed (early response) window or dotted (late response) window. (While it is true that the late responses to the stimulus at time point $t - \Delta t$ overlap with early responses to the stimulus at time point t , the decoder used to reconstruct the stimulus at time point t from early responses is only minimally affected by late responses to the stimulus at time point $t - \Delta t$ when the decoder is estimated by averaging over a long enough duration, e.g., tens of seconds). The cutoff time between early and late responses, $\tau_{boundary}$, was chosen to minimize the overlap between the M50_{TRF} and M100_{TRF} peaks, on a per-subject basis, with a median value of 85 ms (range, 70–100 ms in 5 ms increments); repeating the analysis using the single value of 85 ms for all subjects did not qualitatively change any conclusions. When decoding from early responses only, the time window of integration is from $\tau_b = 0$ to $\tau_e = \tau_{boundary}$. When decoding from late neural responses only, the time window of integration is from $\tau_b = \tau_{boundary}$ to $\tau_e = 500$ ms.

The robustness of different representations, such as of foreground versus background, can be compared by examining the accuracy of their respective stimulus envelope reconstructions (Fig. 1, right).

Statistics. All statistical comparisons reported here are two-tailed permutation tests with $N = 1,000,000$ random permutations (within subject). Due to the value of N selected, the smallest accurate p value that can be reported is $2 \times 1/N (= 2 \times 10^{-6})$; the factor of 2 arises from the two-tailed test) and any p value $< 2/N$ is reported as $p < 2 \times 10^{-6}$. The statistical comparison between foreground and individual backgrounds requires special mention, since each listening condition has one foreground but two individual backgrounds. From the perspective of both behavior and task, both the individual backgrounds are interchangeable. Hence, when comparing reconstruction accuracy of foreground versus individual background, the average reconstruction accuracy of the two individual backgrounds is used. Finally, Bayes factor analysis is used, when appropriate, to evaluate evidence in favor of null hypothesis, since conventional hypothesis testing is not suitable for such purposes. Briefly, Bayes factor analysis calculates the posterior odds, i.e., the ratio of $P(H_0|observations)$ to $P(H_1|observations)$, where H_0 and H_1 are the null and alternate hypotheses respectively (Eqs. 5 and 6):

$$\frac{P(H_0|observations)}{P(H_1|observations)} = \frac{P(observations|H_0)}{P(observations|H_1)} \times \frac{P(H_0)}{P(H_1)}$$

$$= BF_{01} \times \frac{P(H_0)}{P(H_1)}$$

The ratio of $P(observations|H_0)$ and $P(observations|H_1)$ is denoted as the Bayes factor, BF_{01} . Then, under the assumption of equal priors, $P(H_0) = P(H_1)$, the posterior odds reduces to BF_{01} . A BF_{01} value of 10 indicates that the data is 10 times more likely to occur under the null hypothesis than the alternate hypothesis; conversely, a BF_{01} value of 1/10 indicates that the data is 10 times more likely to occur under the alternate hypothesis than the null hypothesis. Conventionally, a BF_{01} value between 3 and 10 is considered as moderate evidence in favor of the null hypothesis, and a value between 10 and 30 is considered strong evidence in favor of the null hypothesis; conversely, a BF_{01} value between 1/3 and 1/10 (respectively 1/10 and 1/30) is considered moderate (respectively strong) evidence for the alternate hypothesis (Rouder et al., 2009).

Results

Stimulus reconstruction from early neural responses

To investigate the neural representations of the attended versus unattended speech streams associated with early auditory areas (i.e., from the core auditory cortex; Nourski et al., 2014), the temporal envelope of attended (foreground) and unattended speech streams (individual backgrounds) were reconstructed using decoders optimized individually for each speech stream. All reconstructions performed significantly better than chance level (foreground vs noise, $p < 2 \times 10^{-6}$; individual background vs noise, $p < 2 \times 10^{-6}$), indicating that all three speech streams are represented in the early auditory cortex. Figure 3A shows reconstruction accuracy for foreground versus individual backgrounds. A permutation test shows no significant difference between foreground and individual background ($p = 0.21$), indicating that there is no evidence of significant neural bias for the attended speech stream over the ignored speech stream, in early neural responses. In fact, Bayes factor analysis ($BF_{01} = 4.2$) indicates moderate support in favor of the null hypothesis (Rouder et al., 2009), which supports the idea that early neural responses do not distinguish significantly between attended and ignored speech streams.

To test the hypothesis that early auditory areas represent the auditory scene in terms of acoustics, rather than as individual auditory objects, we reconstructed the acoustic scene (the envelope of the sum of all three speech streams) and compared it against the reconstruction of the sum of streams (sum of envelopes of each of the three individual speech streams). Separate decoders optimized individually were used to reconstruct the acoustic scene and the sum of streams. As can be seen in Figure 3B, the

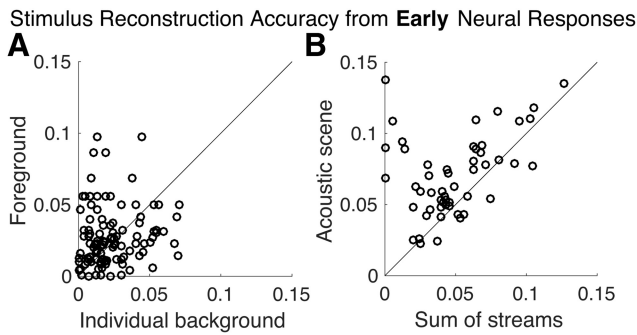


Figure 3. Stimulus envelope reconstruction accuracy using early neural responses. **A**, Scatter plot of reconstruction accuracy of the foreground versus individual background envelopes. No significant difference was observed ($p = 0.21$), and therefore no preferential representation of the foreground stream over the individual background streams is revealed in early neural responses. Each data point corresponds to a distinct background and condition partition per subject (with two backgrounds sharing a common foreground). **B**, Scatter plot of reconstruction accuracy of the envelope of the entire acoustic scene versus that of the sum of the envelopes of all three individual speech streams. The acoustic scene is reconstructed more accurately (visually, most of data points fall above the diagonal) as a whole than as the sum of individual components in early neural responses ($p < 2 \times 10^{-6}$). Each data point corresponds to a distinct condition partition per subject. In both plots, reconstruction accuracy is measured by proportion of the variance explained: the square of the Pearson correlation coefficient between the actual and predicted envelopes.

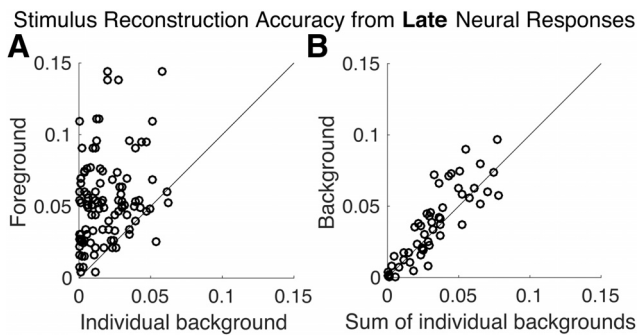


Figure 4. Stimulus envelope reconstruction accuracy using late neural responses. **A**, Scatter plot of accuracy of foreground versus individual background envelope reconstructions demonstrates that the foreground is represented with dramatically better fidelity (visually, most of data points fall above the diagonal) than the background speech, in late neural responses ($p < 2 \times 10^{-6}$). Each data point corresponds to a distinct background and condition partition per subject (with two backgrounds sharing a common foreground). **B**, Scatter plot of the reconstruction accuracy of the envelope of the entire background versus that of the sum of the envelopes of the two individual background speech streams. The background scene is reconstructed more accurately as a monolithic background than as separated individual background streams in late neural responses ($p = 0.012$). Each data point corresponds to a distinct condition partition per subject.

result shows that the acoustic scene is better reconstructed than the sum of streams ($p < 2 \times 10^{-6}$). This indicates that the early auditory cortex is better described as processing the entire acoustic scene rather than processing the separate elements of the scene individually.

Stimulus reconstruction from late neural responses

While the preceding results were based on early cortical processing, the following results are based on late auditory cortical processing (responses with latencies $> \sim 85$ ms). Figure 4A shows the scatter plot of reconstruction accuracy for the foreground versus individual background envelopes based on late responses. A paired permutation test shows that reconstruction accuracy for the foreground is significantly higher than the background

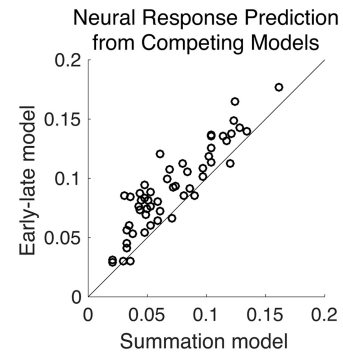


Figure 5. MEG response prediction accuracy. Scatter plot of the accuracy of predicted MEG neural response for the proposed early-late model versus the standard summation model. The early-late model predicts the MEG neural response dramatically better (visually, most of data points fall above the diagonal) than the summation model ($p < 2 \times 10^{-6}$). The accuracy of predicted MEG neural responses is measured by proportion of the variance explained: the square of the Pearson correlation coefficient between the actual and predicted responses. Each data point corresponds to a distinct condition partition per subject.

($p < 2 \times 10^{-6}$). Even though the individual backgrounds are not as reliably reconstructed as the foreground, their reconstructions are nonetheless significantly better than chance level ($p < 2 \times 10^{-6}$).

To distinguish among possible neural representations of the background streams, we compared the reconstructability of the envelope of the entire background as a whole, with the reconstructability of the sum of the envelopes of the (two) backgrounds. If the background is represented as a single auditory object (i.e., “the background”), the reconstruction of the envelope of the entire background should be more faithful than the sum of envelopes of individual backgrounds. In contrast, if the background is represented as distinct auditory objects, each distinguished by its own envelope, the reconstruction of the sum of envelopes of the individual backgrounds should be more faithful. Figure 4B shows the scatter plot of reconstruction accuracy for the envelope of combined background versus the sum of the envelopes of the individual background streams. Analysis shows that the envelope of the combined background is significantly better represented than the sum of the envelopes of the individual backgrounds ($p = 0.012$). As noted previously, the envelope of the combined background is actually strongly correlated with the sum of the envelopes of the individual backgrounds, meaning that finding a significant difference in their reconstruction accuracy is a priori unlikely, providing even more credence to the result.

Encoding analysis

Results above from envelope reconstruction suggest that while early neural responses represent the auditory scene in terms of the acoustics, the later neural responses represent the auditory scene in terms of a separated foreground and a single background stream. To further test this hypothesis, we use TRF-based encoding analysis to directly compare two different models of auditory scene representations. The two models compared are the standard summation model (based on parallel representations of all speech streams; Eq. 2) and the new early-late model (based on an early representation of the entire acoustic scene and late representations of separated foreground and background; Eq. 3). Figure 5 shows the response prediction accuracies for the two models. A permutation test shows that the accuracy of the early-late model is considerably higher than that of the summation

model ($p < 2 \times 10^{-6}$). This indicates that a model in which the early/core auditory cortex processes the entire acoustic scene but the later/higher-order auditory cortex processes the foreground and background separately has more support than the previously employed model of parallel processing of separate streams throughout the auditory cortex.

Discussion

In this study, we used cortical tracking of continuous speech, in a multitalker scenario, to investigate the neural representations of an auditory scene. From MEG recordings of subjects selectively attending to one of three collocated speech streams, we observed the following: (1) the early neural responses (from sources with short latencies), which originate primarily from the core auditory cortex, represent the foreground (attended) and background (ignored) speech streams without any significant difference, whereas the late neural responses (from sources with longer latencies), which originate primarily from higher-order areas of the auditory cortex, represent the foreground with significantly higher fidelity than the background; (2) early neural responses are not only balanced in how they represent the constituent speech streams, but in fact represent the entire acoustic scene holistically, rather than as separately contributing individual perceptual objects; and (3) even though there are two physical speech streams in the background, no neural segregation is observed for the background speech streams.

It is well established that auditory processing in the cortex is performed in a hierarchical fashion, in which an auditory stimulus is processed by different anatomical areas at different latencies (Inui et al., 2006; Nourski et al., 2014). Using this idea to inform the neural decoding/encoding analysis allows the effective isolation of neural signals from a particular cortical area, and thereby the ability to track changes in neural representations as the stimulus processing proceeds along the auditory hierarchy. This time-constrained reconstruction/prediction approach may prove especially fruitful in high-time-resolution/low-spatial-resolution imaging techniques, such as MEG and EEG. Even though different response components are generated by different neural sources, standard neural source localization algorithms may perform poorly when different sources are strongly correlated in their responses (Lutkenhoner and Mosher, 2007). While the proposed method is not to be viewed as an alternative to source localization methods, it can nonetheless be used to tease apart different components of an MEG/EEG response, without explicit source localization.

Even though there is no significant difference between the ability to reconstruct the foreground and background from early neural responses, nonetheless we observe a nonsignificant tendency toward an enhanced representation of the foreground (foreground > background, $p = 0.21$). This could be due to task-related plasticity of spectrotemporal receptive fields of neurons in the mammalian primary auditory cortex (Fritz et al., 2003), where the receptive fields of neurons are tuned to match the stimulus characteristics of attended sounds. The selective amplification of foreground in late neural responses (from higher-order auditory cortices) but not in early responses (from the core auditory cortex) observed here using decoding is in agreement with the encoding result of Ding and Simon (2012b), where the authors showed that the late M100_{TRF} component, but not the early M50_{TRF} component, of TRF is significantly modulated by attention. The increase in fidelity of the foreground as the response latency increases indicates a temporal as well as functional hierarchy in cortical processing of the auditory scene, from core to higher-order areas in the auditory cortex. Similar preferential

representation for the attended speech stream has been demonstrated, albeit with only two speech streams and not differentiating between early and late responses, using δ -band and θ -band neural responses (Ding and Simon, 2012b; Zion Golumbic et al., 2013a,b), as well as high- γ neural responses (Mesgarani and Chang, 2012; Zion Golumbic et al., 2013a), and also for the cases of monaural speech (Ding and Simon, 2012b; Mesgarani and Chang, 2012) and audiovisual speech (Zion Golumbic et al., 2013a,b).

While some researchers suggest selective entrainment (Schröder and Lakatos, 2009; Ng et al., 2012; Zion Golumbic et al., 2013b; Kayser et al., 2015) as the mechanism for selective tracking of attended speech, others suggest a temporal coherence model (Shamma et al., 2011; Ding and Simon, 2012b). Natural speech is quasirhythmic with dominant rates at syllabic, word, and prosodic frequencies. The selective entrainment model suggests that attention causes endogenous low-frequency neural oscillations to align with the temporal structure of the attended speech stream, thus aligning the high-excitability phases of oscillations with events in attended stream. This effectively forms a mask that favors the attended speech. The temporal coherence model suggests that selective tracking of attended speech is achieved in two stages. First, a cortical filtering stage, where feature-selective neurons filter the stimulus, producing a multidimensional representation of the auditory scene along different feature axes. This is followed by a second stage, coherence analysis, which combines relevant feature streams based on their temporal similarity, giving rise to separate perceptions of attended and ignored streams. In this model, it is hypothesized that attention, acting through in the coherence analysis stage, plays an important role in stream formation. This type of coherence model predicts an unsegregated representation of any (nonattended) background streams.

The representation of an auditory scene in the core auditory cortex, based on the early responses, is here shown to be more spectrotemporal or acoustic based than object based (Fig. 3B). This is further supported by the result that the early-late model predicts MEG neural responses significantly better than the summation model (Fig. 5). This is consistent with previous demonstrations showing that neural activity in the core auditory cortex is highly sensitive to acoustic characteristics of speech and primarily reflects spectrotemporal attributes of sound (Nourski et al., 2009; Okada et al., 2010; Ding and Simon, 2013; Steinschneider et al., 2014). In contrast, Nelken and Bar-Yosef (2008) suggest that neural auditory objects may form as early as the primary auditory cortex, and Fritz et al. (2003) show that representations of dynamic sounds in the primary auditory cortex are influenced by task. As a working principle, it is possible that less complex stimuli are resolved earlier in the hierarchy of the auditory pathway (e.g., sounds that can be separated via tonotopy), whereas more complex stimuli (e.g., concurrent speech streams), which need further processing, are resolved only much later in the auditory pathway. In addition, it is worth noting that the current study uses collocated speech streams, whereas mechanisms of stream segregation will also be influenced by other auditory cues, including spatial cues, differences in acoustic source statistics (e.g., only speech streams vs mixed speech and music; strong statistical differences might drive stream segregation in a more bottom-up manner than the top-down attentional effects studied here), and perceptual load effects (e.g., tone streams vs speech streams), as well as visual cues. Any of these additional cues has the potential to alter the timing and neural mechanisms by which auditory scene analysis occurs.

It is widely accepted that an auditory scene is perceived in terms of auditory objects (Bregman, 1994; Griffiths and Warren, 2004; Shinn-Cunningham, 2008; Shamma et al., 2011). Ding and Simon (2012a) demonstrated evidence for an object-based cortical representation of an auditory scene, but did not distinguish between early and late neural responses. This, coupled with the result here that early neural responses provide an acoustic, not object-based, representation, strongly suggests that the object-based representation emerges only in the late neural responses/higher-order (belt and parabelt) auditory areas. This is further supported by the observation that acoustic invariance, a property of object-based representation, is observed in higher-order areas but not in the core auditory cortex (Chang et al., 2010; Okada et al., 2010). When the foreground is represented as an auditory object in late neural responses, the finding that the combined background is better reconstructed than the sum of envelopes of individual backgrounds (Fig. 4B) suggests that in late neural responses the background is not represented as separated and distinct auditory objects. This result is consistent with that of Sussman et al. (2005), who reported an unsegregated background when subjects attended to one of three tone streams in the auditory scene. This unsegregated background may be a result of an “analysis-by-synthesis” (Yuille and Kersten, 2006; Poeppel et al., 2008) mechanism, wherein the auditory scene is first decomposed into basic acoustic elements, followed by top-down processes that guide the synthesis of the relevant components into a single stream, which then becomes the object of attention. The remainder of the auditory scene would be the unsegregated background, which itself might have the properties of an auditory object. When attention shifts, new auditory objects are correspondingly formed, with the old ones now contributing to the unstructured background. Shamma et al. (2011) suggest that this top-down influence acts via the principle of temporal coherence. Between the two opposing views—that streams are formed preattentively and that multiple streams can coexist simultaneously; or that attention is required to form a stream and only that single stream is ever present as a separated perceptual entity—these findings lend support to the latter.

In summary, these results provide evidence that, in a complex auditory scene with multiple overlapping spectral and temporal sources, the core areas of the auditory cortex maintain an acoustic representation of the auditory scene with no significant preference to attended over ignored sources, and with no separation into distinct sources. It is only the higher-order auditory areas that provide an object-based representation for the foreground, but even there the background remains unsegregated.

References

- Ahveninen J, Hämäläinen M, Jääskeläinen IP, Ahlfors SP, Huang S, Lin FH, Raij T, Sams M, Vasios CE, Belliveau JW (2011) Attention-driven auditory cortex short-term plasticity helps segregate relevant sounds from noise. *Proc Natl Acad Sci U S A* 108:4182–4187. [CrossRef Medline](#)
- Bregman AS (1994) Auditory scene analysis: the perceptual organization of sound. Cambridge, MA: MIT.
- Carlyon RP (2004) How the brain separates sounds. *Trends Cogn Sci* 8:465–471. [CrossRef Medline](#)
- Chang EF, Rieger JW, Johnson K, Berger MS, Barbaro NM, Knight RT (2010) Categorical speech representation in human superior temporal gyrus. *Nat Neurosci* 13:1428–1432. [CrossRef Medline](#)
- Cherry EC (1953) Some experiments on the recognition of speech, with one and with 2 ears. *J Acoust Soc Am* 25:975–979. [CrossRef](#)
- Chi T, Ru P, Shamma SA (2005) Multiresolution spectrotemporal analysis of complex sounds. *J Acoust Soc Am* 118:887–906. [CrossRef Medline](#)
- David SV, Mesgarani N, Shamma SA (2007) Estimating sparse spectrotemporal receptive fields with natural stimuli. *Network* 18:191–212. [CrossRef Medline](#)
- Davis MH, Johnsrude IS (2003) Hierarchical processing in spoken language comprehension. *J Neurosci* 23:3423–3431. [Medline](#)
- de Cheveigné A, Parra LC (2014) Joint decorrelation, a versatile tool for multichannel data analysis. *Neuroimage* 98:487–505. [CrossRef Medline](#)
- de Cheveigné A, Simon JZ (2007) Denoising based on time-shift PCA. *J Neurosci Methods* 165:297–305. [CrossRef Medline](#)
- de Cheveigné A, Simon JZ (2008) Denoising based on spatial filtering. *J Neurosci Methods* 171:331–339. [CrossRef Medline](#)
- Di Liberto GM, O’Sullivan JA, Lalor EC (2015) Low-frequency cortical entrainment to speech reflects phoneme-level processing. *Curr Biol* 25:2457–2465. [CrossRef Medline](#)
- Ding N, Simon JZ (2012a) Neural coding of continuous speech in auditory cortex during monaural and dichotic listening. *J Neurophysiol* 107:78–89. [CrossRef Medline](#)
- Ding N, Simon JZ (2012b) Emergence of neural encoding of auditory objects while listening to competing speakers. *Proc Natl Acad Sci U S A* 109:11854–11859. [CrossRef Medline](#)
- Ding N, Simon JZ (2013) Adaptive temporal encoding leads to a background-insensitive cortical representation of speech. *J Neurosci* 33:5728–5735. [CrossRef Medline](#)
- Fritz J, Shamma S, Elhilali M, Klein D (2003) Rapid task-related plasticity of spectrotemporal receptive fields in primary auditory cortex. *Nat Neurosci* 6:1216–1223. [CrossRef Medline](#)
- Griffiths TD, Warren JD (2002) The planum temporale as a computational hub. *Trends Neurosci* 25:348–353. [CrossRef Medline](#)
- Griffiths TD, Warren JD (2004) What is an auditory object? *Nat Rev Neurosci* 5:887–892. [CrossRef Medline](#)
- Hickok G, Poeppel D (2007) The cortical organization of speech processing. *Nat Rev Neurosci* 8:393–402. [CrossRef Medline](#)
- Inui K, Okamoto H, Miki K, Gunji A, Kakigi R (2006) Serial and parallel processing in the human auditory cortex: a magnetoencephalographic study. *Cereb Cortex* 16:18–30. [CrossRef Medline](#)
- Kaas JH, Hackett TA (2000) Subdivisions of auditory cortex and processing streams in primates. *Proc Natl Acad Sci U S A* 97:11793–11799. [CrossRef Medline](#)
- Kayser C, Wilson C, Safaai H, Sakata S, Panzeri S (2015) Rhythmic auditory cortex activity at multiple timescales shapes stimulus-response gain and background firing. *J Neurosci* 35:7750–7762. [CrossRef 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](#)
- Lalor EC, Power AJ, Reilly RB, Foxe JJ (2009) Resolving precise temporal processing properties of the auditory system using continuous stimuli. *J Neurophysiol* 102:349–359. [CrossRef Medline](#)
- Lutkenhoner B, Mosher JC (2007) Source analysis of auditory evoked potentials and fields. In: Auditory evoked potentials: basic principles and clinical application (Burkard RF, Eggermont JJ, Don M, eds). Philadelphia: Lippincott Williams and Wilkins.
- McDermott JH (2009) The cocktail party problem. *Curr Biol* 19:R1024–R1027. [CrossRef Medline](#)
- Mesgarani N, Chang EF (2012) Selective cortical representation of attended speaker in multi-talker speech perception. *Nature* 485:233–236. [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](#)
- Nelken I, Bar-Yosef O (2008) Neurons and objects: the case of auditory cortex. *Front Neurosci* 2:107–113. [CrossRef Medline](#)
- Ng BS, Schroeder T, Kayser C (2012) A precluding but not ensuring role of entrained low-frequency oscillations for auditory perception. *J Neurosci* 32:12268–12276. [CrossRef Medline](#)
- Nourski KV, Reale RA, Oya H, Kawasaki H, Kovach CK, Chen H, Howard MA 3rd, Brugge JF (2009) Temporal envelope of time-compressed speech represented in the human auditory cortex. *J Neurosci* 29:15564–15574. [CrossRef Medline](#)
- Nourski KV, Steinschneider M, McMurray B, Kovach CK, Oya H, Kawasaki H, Howard MA 3rd (2014) Functional organization of human auditory cortex: investigation of response latencies through direct recordings. *Neuroimage* 101:598–609. [CrossRef Medline](#)
- Okada K, Rong F, Venezia J, Matchin W, Hsieh IH, Saberi K, Serences JT,

- Hickok G (2010) Hierarchical organization of human auditory cortex: evidence from acoustic invariance in the response to intelligible speech. *Cereb Cortex* 20:2486–2495. [CrossRef Medline](#)
- Okamoto H, Stracke H, Bermudez P, Pantev C (2011) Sound processing hierarchy within human auditory cortex. *J Cogn Neurosci* 23:1855–1863. [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](#)
- Overath T, McDermott JH, Zarate JM, Poeppel D (2015) The cortical analysis of speech-specific temporal structure revealed by responses to sound quilts. *Nat Neurosci* 18:903–911. [CrossRef Medline](#)
- Pasley BN, David SV, Mesgarani N, Flinker A, Shamma SA, Crone NE, Knight RT, Chang EF (2012) Reconstructing speech from human auditory cortex. *PLoS Biol* 10:e1001251. [CrossRef Medline](#)
- Peelle JE, Johnsrude IS, Davis MH (2010) Hierarchical processing for speech in human auditory cortex and beyond. *Front Hum Neurosci* 4:51. [CrossRef Medline](#)
- Poeppel D, Idsardi WJ, van Wassenhove V (2008) Speech perception at the interface of neurobiology and linguistics. *Philos Trans R Soc Lond B Biol Sci* 363:1071–1086. [CrossRef Medline](#)
- Power AJ, Foxe JJ, Forde EJ, Reilly RB, Lalor EC (2012) At what time is the cocktail party? A late locus of selective attention to natural speech. *Eur J Neurosci* 35:1497–1503. [CrossRef Medline](#)
- Rauschecker JP, Scott SK (2009) Maps and streams in the auditory cortex: nonhuman primates illuminate human speech processing. *Nat Neurosci* 12:718–724. [CrossRef Medline](#)
- Recanzone GH, Guard DC, Phan ML (2000) Frequency and intensity response properties of single neurons in the auditory cortex of the behaving macaque monkey. *J Neurophysiol* 83:2315–2331. [Medline](#)
- Rouder JN, Speckman PL, Sun D, Morey RD, Iverson G (2009) Bayesian t tests for accepting and rejecting the null hypothesis. *Psychon Bull Rev* 16:225–237. [CrossRef Medline](#)
- Särelä J, Valpola H (2005) Denoising source separation. *J Machine Learn Res* 6:233–272. [CrossRef](#)
- Schroeder CE, Lakatos P (2009) Low-frequency neuronal oscillations as instruments of sensory selection. *Trends Neurosci* 32:9–18. [CrossRef Medline](#)
- Shamma SA, Elhilali M, Micheyl C (2011) Temporal coherence and attention in auditory scene analysis. *Trends Neurosci* 34:114–123. [CrossRef Medline](#)
- Shinn-Cunningham BG (2008) Object-based auditory and visual attention. *Trends Cogn Sci* 12:182–186. [CrossRef Medline](#)
- Steinschneider M, Liégeois-Chauvel C, Brugge JF (2011) Auditory evoked potentials and their utility in the assessment of complex sound processing. In: *The auditory cortex*, pp 535–559. New York: Springer.
- Steinschneider M, Nourski KV, Rhone AE, Kawasaki H, Oya H, Howard MA 3rd (2014) Differential activation of human core, non-core and auditory-related cortex during speech categorization tasks as revealed by intracranial recordings. *Front Neurosci* 8:240. [CrossRef Medline](#)
- Sussman ES, Bregman AS, Wang WJ, Khan FJ (2005) Attentional modulation of electrophysiological activity in auditory cortex for unattended sounds within multistream auditory environments. *Cogn Affect Behav Neurosci* 5:93–110. [CrossRef Medline](#)
- Sweet RA, Dorph-Petersen KA, Lewis DA (2005) Mapping auditory core, lateral belt, and parabelt cortices in the human superior temporal gyrus. *J Comp Neurol* 491:270–289. [CrossRef Medline](#)
- Yuille A, Kersten D (2006) Vision as Bayesian inference: analysis by synthesis? *Trends Cogn Sci* 10:301–308. [CrossRef Medline](#)
- Zion Golumbic E, Cogan GB, Schroeder CE, Poeppel D (2013a) Visual input enhances selective speech envelope tracking in auditory cortex at a “cocktail party.” *J Neurosci* 33:1417–1426. [CrossRef Medline](#)
- Zion Golumbic EM, Ding N, Bickel S, Lakatos P, Schevon CA, McKhann GM, Goodman RR, Emerson R, Mehta AD, Simon JZ, Poeppel D, Schroeder CE (2013b) Mechanisms underlying selective neuronal tracking of attended speech at a “cocktail party.” *Neuron* 77:980–991. [CrossRef Medline](#)