Cortical Representations of Speech in a Multi-talker Auditory Scene

Krishna C. Puvvada[a], Jonathan Z. Simon[a,b,c]

[a] Department of Electrical & Computer Engineering, University of Maryland,
College Park, MD 20742

[b] Department of Biology, University of Maryland, College Park, MD 20742

[c] Institute for Systems Research, University of Maryland, College Park, MD 20742

Corresponding author: Jonathan Z. Simon (jzsimon@umd.edu), Department of Electrical & Computer Engineering, University of Maryland, 8223 Paint Branch Dr., College Park, MD 20742.

Abbreviated title: **Foreground and Background Speech Representations**

Number of pages: 42

Number of figures: 5

Number of words in Abstract: 213

Number of words in Introduction: 621

Number of words in Discussion: 1485

Conflict of Interest: The authors declare no competing financial interests.

Acknowledgements: The authors thank Elizabeth Nguyen for excellent MEG technical support, and Stuart Rosen for advice on the stimulus paradigm. This work was supported by NIH grant R01 DC014085.

Abstract

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-objects based representation in auditory cortex. Here, using magnetoencephalography (MEG) recordings from human subjects, we investigate how a complex acoustic scene consisting of multiple speech sources is represented in distinct hierarchical stages of auditory cortex. Using systems-theoretic methods of stimulus reconstruction, we show that the primary-like areas in auditory cortex contain dominantly spectro-temporal 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. In contrast, we also show that higher order auditory cortical areas 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. Taken together, these findings demonstrate the progression of the representations and processing of a complex acoustic scene up through the hierarchy of human auditory cortex.

Significance Statement:

Using magnetoencephalography (MEG) 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 auditory cortex. We show that the primary-like areas in auditory cortex use a dominantly spectro-temporal based representation of the entire auditory scene, with both attended and ignored speech streams represented with almost equal fidelity. In contrast, we show that higher order auditory cortical areas 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 question, yet the brain nevertheless routinely performs this task with ease. The neural mechanisms by which this perceptual 'un-mixing' 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 primate cortex is hierarchical (Davis and Johnsrude, 2003; Hickok and Poeppel, 2007; Rauschecker and Scott, 2009; Okada et al., 2010; Peelle et al., 2010) with subcortical areas projecting onto the core areas of auditory cortex, and from there, on to belt, parabelt and additional auditory areas (Kaas and Hackett, 2000). Sound entering the ear reaches different anatomical/functional areas of 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 non-invasive 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$_{\mathrm{TRF}}$ and M100$_{\mathrm{TRF}}$, with respective latencies of 30 – 80 ms and 80 – 150 ms, of which the dominant neural sources are, respectively, Heschl's gyrus (HG) and Planum temporale (PT) (Steinschneider et al., 2011; Ding and Simon, 2012a). Posteromedial HG is the site of core auditory cortex; 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 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 if 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 (Ding and Simon, 2012a; Mesgarani and Chang, 2012; Zion Golumbic et al., 2013b). This results in a limitation, which cannot address the question of distinct vs. collective neural representations for unattended sources. This touches on the long-standing debate of whether auditory object segregation is pre-attentive or it 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 multi-talker auditory scene. We test two major hypotheses: that the dominant representation in 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 (Ding and Simon, 2012a; Mesgarani and Chang, 2012; Di Liberto et al., 2015).

Materials & Methods:

Subjects & Experimental Design Nine normal-hearing, young adults (6 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, female adult and a child speaker. 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 & Wilhelm Grimm (https://librivox.org/fairy-tales-by-the-brothers-grimm/). Periods of silence longer than 300 ms were replaced by a shorter gap whose duration was chosen randomly between 200 ms and 300 ms. The audio signal was low-pass filtered below 4 kHz. In first of three conditions, 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 two, during which the same mixture was played as in condition one, the subjects were instead asked to attend to the male adult speaker (with female adult and child speakers as background). Similarly, in condition three, the target was switched to the female adult speaker. Each condition was repeated three times successively, producing three trials per condition. The presentation order of the three conditions was counterbalanced across subjects. Each trial was of 220 s duration, divided into two 110 s sections, to reduce listener fatigue. To help participants attend to the correct speaker, the first 30 s of each section 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 segment 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 recount a short summary of 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. This data was used check whether any potential subjects gave abnormal auditory responses, but no subjects were excluded based on this criterion.

Data recording and pre-processing 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/$\sqrt{\text{Hz}}$ or better in the white noise region. Subjects lay horizontally in a dimly lit magnetically shielded room (Yokogawa Electric Corporation). Responses were recorded with a sampling rate of 1 kHz with an online 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 utilized to reduce environmental noise from the MEG recordings using the Time-Shift PCA method (de Cheveigne and Simon, 2007). Additionally, MEG recordings were decomposed into virtual sensors/ components using denoising source separation (DSS) (Särelä and Valpola, 2005; de Cheveigne and Simon, 2008; de Cheveigne 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, for all further analysis the 157 MEG sensors were reduced, using DSS, to 4 components in each hemisphere. Also, both stimulus envelope and MEG responses were band pass filtered between 1 – 8 Hz (delta and theta bands), which correspond to the slow temporal modulations in speech (Ding and Simon, 2012b, a).

Terminology and Notation As specified in the stimulus description, in each condition the subject attends to one among the three speech streams. The envelope of 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 entire acoustic stimulus or auditory scene, comprising of all the 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’, and the two are not mathematically equal: even though both are functions of the same stimuli, they differ due to the non-linear nature of a signal envelope (the linear correlation between the acoustic scene and the sum of streams is typically ~0.75).

Neural responses with latencies less than ~85 ms (typically originating from core auditory areas) are referred to here as ‘early neural responses’ and responses with latencies more than ~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’.

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

$$r(t) = \sum_{\tau} s(t-\tau)TRF(\tau) + \varepsilon(t) \qquad (1)$$

where $t = 0,1, \ldots, T$ 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, 2012b). 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 ms and 80 ms and a negative peak between 90 ms and 130 ms, referred to as M50$_{\mathrm{TRF}}$ and M100$_{\mathrm{TRF}}$ respectively (Ding and Simon, 2012a) (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, 2012a; Zion Golumbic et al., 2013b), referred to here as the 'Summation model'. For example, with two speech streams, the neural response would be modeled as

$$r(t) = \sum_{\tau} S_a(t-\tau)TRF_a(\tau) + \sum_{\tau} S_b(t-\tau)TRF_b(\tau) + \varepsilon(t) \qquad (2)$$

where $S_a(t)$ and $S_b(t)$ are the envelopes of the two speech streams, and $TRF_a(t)$, and $TRF_b(t)$ are the TRFs corresponding to each stream. The summation model is easily extended to the case of more than two speech streams, by adding new terms with each new individual speech stream envelope and the corresponding TRF.

In addition to the existing summation model, we propose a new encoding-model referred to as the ‘Early-late model’, which allows 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.

$$r(t) = \sum_{\tau=0}^{\tau=\tau_1} S_A(t-\tau)TRF_A(\tau) + \sum_{\tau=\tau_1}^{\tau=\tau_2} S_F(t-\tau)TRF_F(\tau) + \sum_{\tau=\tau_1}^{\tau=\tau_2} S_B(t-\tau)TRF_B(\tau) + \epsilon(t) \qquad (3)$$

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 $TRFs. \tau_1, \tau_2$ represent the boundary values of the integration windows for early and late neural responses respectively.

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 (illustrated in Figure 1, left).

(Figure 1 about here)

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

$$E(t) = \sum_{k=1}^{N} \sum_{\tau=\tau_b}^{\tau_e} M_k(t+\tau) D_k(\tau) + \epsilon(t) \tag{4}$$

where $E(t)$ is the reconstructed 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-talker case, the envelope to be reconstructed might be the envelope of the speech of attended talker, or one of the background talkers, or of a mixture 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 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 cut-off time between early and late responses, $\tau_{boundary}$, was chosen to minimize the overlap between the $\mathrm{M50_{TRF}}$ and $\mathrm{M100_{TRF}}$ peaks, on a per subject basis, with a typical value being 85 ms. 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.

(Figure 2 about here)

The robustness of different representations, such as of Foreground vs. Background, can be compared by examining the accuracy of their respective stimulus envelope reconstructions (illustrated in Figure 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 2×1/N (= 2×10^{-6}; the factor of 2 arises from the two-tailed test) and any p value smaller than 2/N is reported as $p < 2\times10^{-6}$. 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 vs. 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.

$$\frac{P(H_0|observations)}{P(H_1|observations)} = \frac{P(observations|H_0)}{P(observations|H_1)} \times \frac{P(H_0)}{P(H_1)} \qquad (5)$$

$$= BF_{01} \times \frac{P(H_0)}{P(H_1)} \qquad (6)$$

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 ten times more likely to occur under the null hypothesis than the alternate hypothesis; conversely, a BF_{01} value of 0.1 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; conversely, a BF_{01} value between 1/3 & 1/10 (respectively 1/10 & 1/30) is considered moderate (respectively strong) evidence for the alternate hypothesis (for more details we refer the reader to Rouder et al. (2009)).

Results

Stimulus reconstruction from early neural responses

To investigate the neural representations of the attended vs. unattended speech streams associated with early auditory areas, i.e., from 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, $p < 2\times10^{-6}$; individual background vs. noise, $p < 2\times10^{-6}$), indicating that all three speech streams are represented in early auditory cortex. Figure 3A shows reconstruction accuracy for foreground vs. 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), that early neural responses do not distinguish significantly between attended and ignored speech streams.

(Figure 3 about here)

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 reconstruction 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 result shows that the acoustic scene is better reconstructed than the sum of streams ($p < 2\times10^{-6}$). This indicates that 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 more than ~85 ms). Figure 4A shows the scatter plot of reconstruction accuracy for the foreground vs. individual background envelopes based on late responses. A paired permutation test shows that reconstruction accuracy for the foreground is significantly higher than the background ($p < 2\times10^{-6}$). Even though the individual backgrounds are not as reliably reconstructed as foreground, their reconstructions are nonetheless significantly better than chance level ($p < 2\times10^{-6}$).

In order 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 vs. 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 individual 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.

(Figure 4 about here)

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. In order 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; see Equation 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; see Equation 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 ($p < 2\times10^{-6}$). This indicates that a model in which early/core auditory cortex processes the entire acoustic scene but 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 auditory cortex.

(Figure 5 about here)

Discussion

In this study, we used cortical tracking of continuous speech, in a multi-talker scenario, to investigate the neural representations of an auditory scene. Differing latencies of the neural sources processing the same stimuli allow us to separate the source activity temporally, thus enabling the tracking of differing neural representations of the auditory scene. From MEG recordings of subjects selectively attending to one of the three co-located speech streams, we observed that 1) The early neural responses (with short latencies), which originate primarily from core auditory cortex, represent the foreground (attended) and background (ignored) speech streams without any significant difference, whereas the late neural responses (with longer latencies), which originate primarily from higher-order areas of 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; 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 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 MEG/EEG response, without explicit source localization.

The envelope reconstruction using the early, auditory core, neural response component showed no significant difference between foreground and background, in contrast to reconstruction using the late, higher-order auditory, neural responses, where the foreground is substantially better represented than any individual background. This *decoding* result is in agreement with the *encoding* result of (Ding and Simon, 2012a) where the authors showed that the early $M50_{TRF}$ component of the temporal response function is not significantly modulated by attention, whereas the late $M100_{TRF}$ component is modulated by attention.

Even though there is no significant difference between the ability to reconstruct the foreground and background from early neural responses, nonetheless we observe a non-significant tendency towards an enhanced representation of the foreground (foreground > background, $p = 0.21$). This could be due to task-related plasticity of spectro-temporal receptive fields of neurons in mammalian primary auditory cortex (Fritz et al., 2003), where the receptive fields of neurons are tuned to match the stimulus characteristics of attended sounds. It could also be explained by entrainment (Schroeder and Lakatos, 2009; Zion Golumbic et al., 2012), which postulates that the high excitability periods of neurons become aligned with temporal structure of foreground, thereby enhancing its neural representation.

The increase in fidelity of the foreground as the response latency increases, from early neural responses (from core auditory cortex) to late neural responses (from higher-order auditory cortex), indicates a temporal as well as functional hierarchy in cortical processing of auditory scene, from core to higher-order areas in auditory cortex. Similar preferential representation for the attended speech stream has been demonstrated, albeit with only two speech streams, using delta and theta band neural responses (Ding and Simon, 2012a; Zion Golumbic et al., 2013a; Zion Golumbic et al., 2013b) as well as high-gamma neural responses (Mesgarani and Chang, 2012; Zion Golumbic et al., 2013a), and using monaural (Ding and Simon, 2012a; Mesgarani and Chang, 2012) as well as audio-visual speech (Zion Golumbic et al., 2013a; Zion Golumbic et al., 2013b).

While some researchers suggest a selective entrainment model (Schroeder and Lakatos, 2009; Zion Golumbic et al., 2013b) as the mechanism underlying the selective tracking of attended speech, others suggest a temporal coherence model (Shamma et al., 2011; Ding and Simon, 2012a) as the neuronal mechanism underlying selective tracking. Natural speech is quasi-rhythmic with different 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 through two stages. First is a cortical filtering stage, where feature selective neurons filter the stimulus producing a multidimensional representation of auditory scene along different feature axes. This is followed by a second stage, coherence analysis, which combines different features streams based on their temporal similarity, giving rise to separate perceptions of attended and ignored streams.

The representation of an auditory scene in core auditory cortex is here shown to be more spectro-temporal- or acoustic-based than object-based, as demonstrated by the result that the envelope of the auditory scene is better reconstructed than the sum of envelopes of the individual speech streams (e.g., Figure 3B). This is further supported by the result that the Early-late model predicts MEG neural responses significantly better than Summation model (e.g., Figure 5). This is consistent with previous studies that demonstrated that neural activity in core auditory cortex was highly sensitive to acoustic characteristics of speech and primarily reflects spectro-temporal attributes of sound (Nourski et al., 2009; Okada et al., 2010; Steinschneider et al., 2014). All these results suggest that early neural responses, primarily from core auditory cortex, reflect an acoustic-based representation rather than object-based. In contrast, Nelken and Bar-Yosef (2008) suggest that neural auditory objects may form as early as primary auditory cortex, and Fritz et al. (2003) show that representations of dynamic sounds in primary auditory cortex are influence by task. It is possible that less complex stimuli are resolved earlier in the hierarchy of auditory pathway (e.g., sounds that can be separated via tonotopy) whereas speech streams, which overlap both spectrally and temporally, are resolved only much later in auditory pathway.

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 (2012b) 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 suggest 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 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 (Figure 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 through the principle of temporal coherence. Between the two opposing views, that streams are formed pre-attentively and that multiple streams can co-exist simultaneously, or that attention is required to form a stream and only that single stream is ever present as 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 auditory cortex maintains an acoustic representation of the auditory scene with no significant preference to attended over ignored source, 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, Hamalainen M, Jaaskelainen 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.

Bregman AS (1994) Auditory scene analysis: The perceptual organization of sound: MIT press.

Carlyon RP (2004) How the brain separates sounds. Trends Cogn Sci 8:465-471.

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.

Cherry EC (1953) Some Experiments on the Recognition of Speech, with One and with 2 Ears. Journal of the Acoustical Society of America 25:975-979.

Chi T, Ru P, Shamma SA (2005) Multiresolution spectrotemporal analysis of complex sounds. J Acoust Soc Am 118:887-906.

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

Davis MH, Johnsrude IS (2003) Hierarchical processing in spoken language comprehension. J Neurosci 23:3423-3431.

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

de Cheveigne A, Simon JZ (2008) Denoising based on spatial filtering. J Neurosci Methods 171:331-339.

de Cheveigne A, Parra LC (2014) Joint decorrelation, a versatile tool for multichannel data analysis. Neuroimage 98:487-505.

Di Liberto GM, O'Sullivan JA, Lalor EC (2015) Low-Frequency Cortical Entrainment to Speech Reflects Phoneme-Level Processing. Curr Biol 25:2457-2465.

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

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

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.

Griffiths TD, Warren JD (2002) The planum temporale as a computational hub. Trends Neurosci 25:348-353.

Griffiths TD, Warren JD (2004) What is an auditory object? Nature Reviews Neuroscience 5:887-892.

Hickok G, Poeppel D (2007) The cortical organization of speech processing. Nat Rev Neurosci 8:393-402.

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.

Kaas JH, Hackett TA (2000) Subdivisions of auditory cortex and processing streams in primates. Proc Natl Acad Sci U S A 97:11793-11799.

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), pp xix, 731 p., 716 p. of plates. Philadelphia: Lippincott Williams & Wilkins.

McDermott JH (2009) The cocktail party problem. Curr Biol 19:R1024-1027.

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

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.

Nelken I, Bar-Yosef O (2008) Neurons and objects: the case of auditory cortex. Front Neurosci 2:107-113.

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.

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.

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.

Okamoto H, Stracke H, Bermudez P, Pantev C (2011) Sound processing hierarchy within human auditory cortex. Journal of Cognitive Neuroscience 23:1855-1863.

Peelle JE, Johnsrude IS, Davis MH (2010) Hierarchical processing for speech in human auditory cortex and beyond. Front Hum Neurosci 4:51.

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.

Rauschecker JP, Scott SK (2009) Maps and streams in the auditory cortex: nonhuman primates illuminate human speech processing. Nat Neurosci 12:718-724.

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.

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.

Särelä J, Valpola H (2005) Denoising source separation. Journal of Machine Learning Research 6:233-272.

Schroeder CE, Lakatos P (2009) Low-frequency neuronal oscillations as instruments of sensory selection. Trends Neurosci 32:9-18.

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

Shinn-Cunningham BG (2008) Object-based auditory and visual attention. Trends Cogn Sci 12:182-186.

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: 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.

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.

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.

Yuille A, Kersten D (2006) Vision as Bayesian inference: analysis by synthesis? Trends in cognitive sciences 10:301-308.

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.

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.

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.

Legend:

Figure 1: Illustrations of different decoding- and encoding-based neural representations of the auditory scene and its constituents. (*Left)* Examples of predicted MEG neural response using the Early-late model (red) and the Summation model (magenta) superimposed on actual MEG response (black). The proposed Early-late model prediction shows higher correlation with the actual MEG neural response than Summation model. (*Right)* Example of speech envelopes reconstructed (grey) from their late neural responses, for both the foreground and the background, superimposed on actual speech envelopes of foreground (blue) and background (cyan). The foreground reconstruction shows higher correlation with the actual foreground envelope, compared to the background reconstruction with the actual background envelope. All examples are grand averages across subjects (3 seconds duration).

Figure 2: Early vs. late MEG neural responses to a continuous speech stimulus. A sample stimulus envelope and multi-channel MEG recordings are shown in red and black respectively. The two grey vertical lines indicate two arbitrary time points at t - Δt and t. The dashed and dotted boxes represent the early and late MEG neural responses to stimulus at time point t respectively. The reconstruction

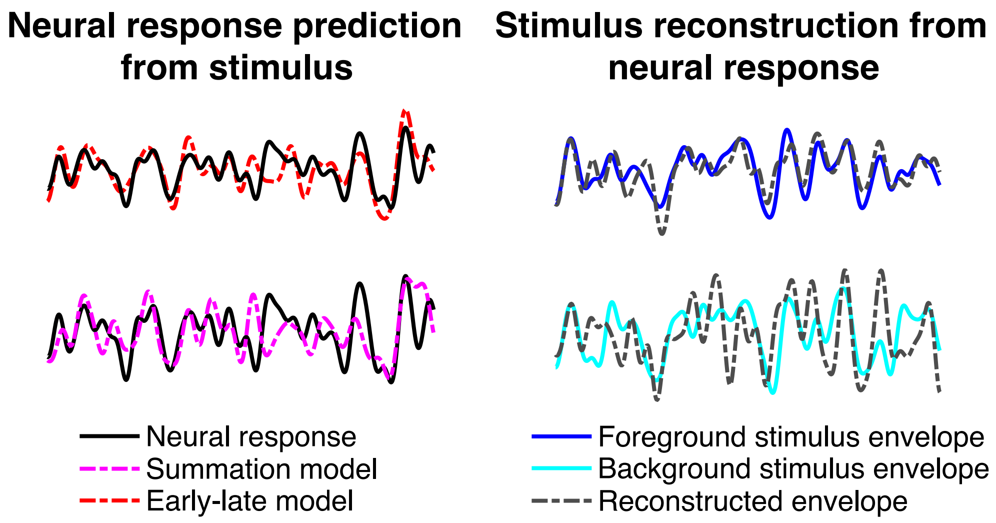
of the stimulus envelope at time *t* can be based on either early or late neural responses, and the separate reconstructions can be compared against each other.

Figure 3: Stimulus envelope reconstruction accuracy using *early* neural responses. **A.** Scatter plot of reconstruction accuracy of the foreground vs. individual background envelopes. No significant difference was observed ($p = 0.21$), and therefore no preferential representation of the foreground speech over the individual background streams is revealed in early neural responses. **B.** Scatter plot of reconstruction accuracy of the envelope of the entire acoustic scene vs. 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 ($p < 2 \times 10^{-6}$). 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 between foreground vs. 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 ($p < 2 \times 10^{-6}$). **B.** Scatter plot of the reconstruction accuracy of the envelope of the entire background vs. 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)

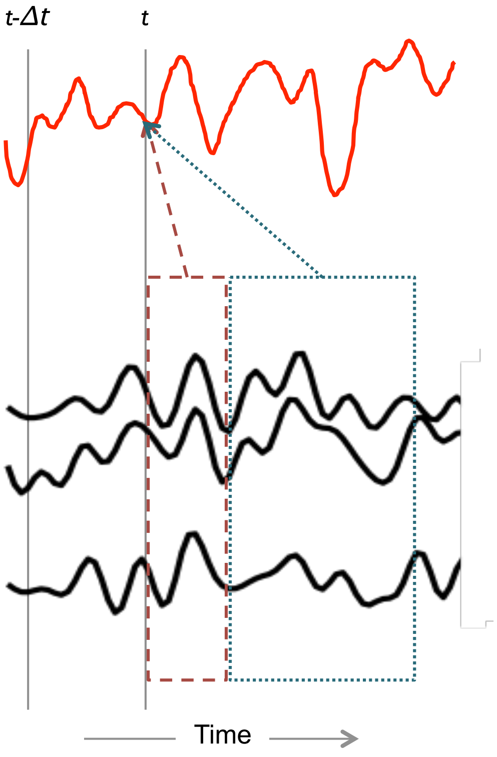
Figure 5: MEG response prediction accuracy. Scatter plot of the accuracy of predicted MEG neural response for the proposed Early-late model vs. the standard Summation model. The Early-late model predicts the MEG neural response dramatically better (visually, most of data points fall above the diagonal) 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.



693

694

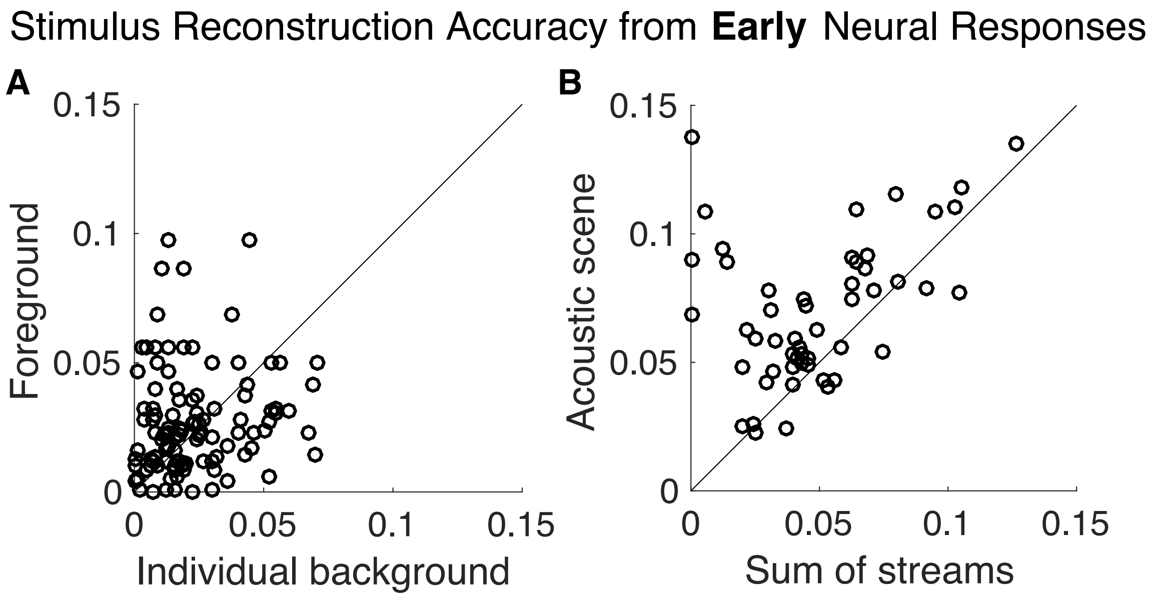
695 Figure 1



696

697

698 Figure 2



Stimulus Reconstruction Accuracy from **Early** Neural Responses

Figure 3

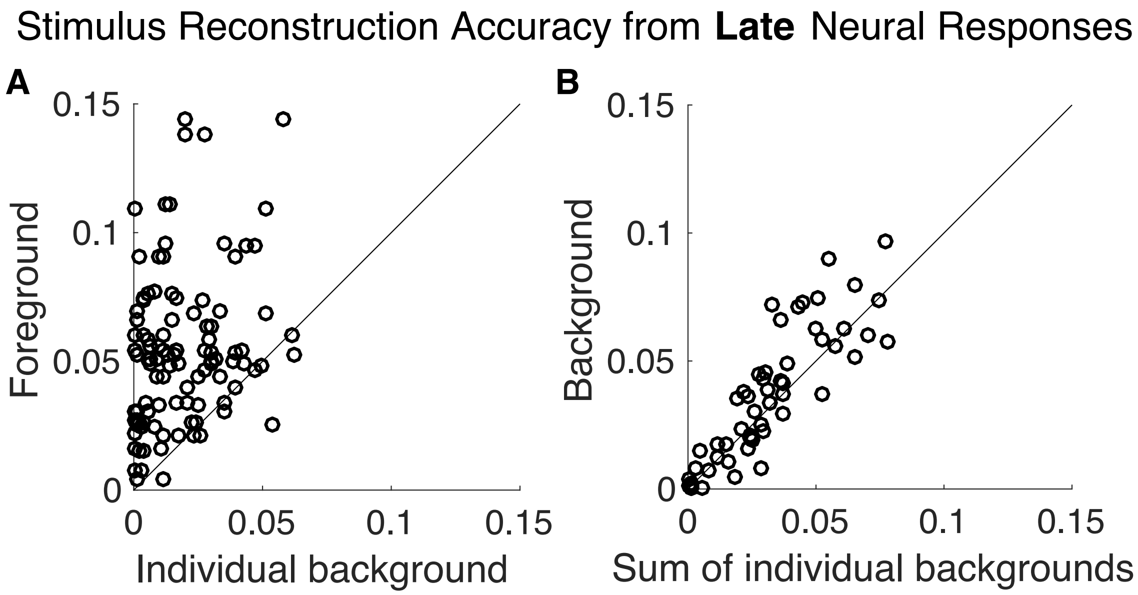

Stimulus Reconstruction Accuracy from **Late** Neural Responses
A
0.15
0.1
0.05
0
Foreground
0
0.05
0.1
0.15
Individual background
B
0.15
0.1
0.05
0
Background
0
0.05
0.1
0.15
Sum of individual backgrounds

Figure 4

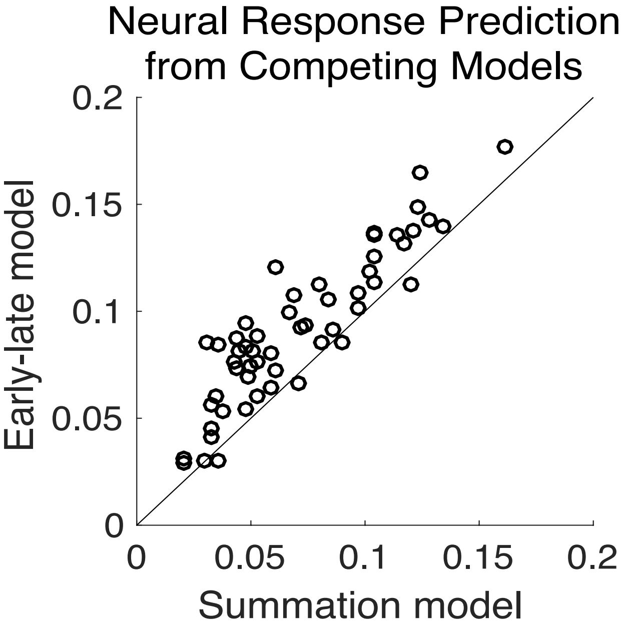

Neural Response Prediction
from Competing Models
0.2
0.15
0.1
0.05
0
Early-late model
0
0.05
0.1
0.15
0.2
Summation model

Figure 5