*J Neurophysiol* 107: 78–89, 2012.
First published October 5, 2011; doi:10.1152/jn.00297.2011.

# Neural coding of continuous speech in auditory cortex during monaural and dichotic listening

**Nai Ding[1] and Jonathan Z. Simon[1,2]**
[1]*Department of Electrical and Computer Engineering and* [2]*Department of Biology, University of Maryland, College Park, Maryland*

Submitted 1 April 2011; accepted in final form 28 September 2011

**Ding N, Simon JZ.** Neural coding of continuous speech in auditory cortex during monaural and dichotic listening. *J Neurophysiol* 107: 78–89, 2012. First published October 5, 2011; doi:10.1152/jn.00297.2011.—The cortical representation of the acoustic features of continuous speech is the foundation of speech perception. In this study, noninvasive magnetoencephalography (MEG) recordings are obtained from human subjects actively listening to spoken narratives, in both simple and cocktail party-like auditory scenes. By modeling how acoustic features of speech are encoded in ongoing MEG activity as a spectrotemporal response function, we demonstrate that the slow temporal modulations of speech in a broad spectral region are represented bilaterally in auditory cortex by a phase-locked temporal code. For speech presented monaurally to either ear, this phase-locked response is always more faithful in the right hemisphere, but with a shorter latency in the hemisphere contralateral to the stimulated ear. When different spoken narratives are presented to each ear simultaneously (dichotic listening), the resulting cortical neural activity precisely encodes the acoustic features of both of the spoken narratives, but slightly weakened and delayed compared with the monaural response. Critically, the early sensory response to the attended speech is considerably stronger than that to the unattended speech, demonstrating top-down attentional gain control. This attentional gain is substantial even during the subjects' very first exposure to the speech mixture and therefore largely independent of knowledge of the speech content. Together, these findings characterize how the spectrotemporal features of speech are encoded in human auditory cortex and establish a single-trial-based paradigm to study the neural basis underlying the cocktail party phenomenon.

speech segregation; attention; spectrotemporal response function; magnetoencephalography

SPOKEN LANGUAGE IS THE DOMINANT form of human communication, and human listeners are superb at tracking and understanding speech even in the presence of interfering speakers (Bronkhorst 2000; Cherry 1953). The critical acoustic features of speech are distributed across several distinct spectral and temporal scales. The slow temporal modulations and coarse spectral modulations reflect the rhythm of speech and contain syllabic and phrasal level segmentation information (Greenberg 1999) and are particularly important for speech intelligibility (Shannon et al. 1995). The neural tracking of slow temporal modulations of speech (e.g., 1–10 Hz) in human auditory cortex can be studied noninvasively using magnetoencephalography (MEG) and electroencephalography (EEG). The low-frequency, large-scale synchronized neural activity recorded by MEG/EEG has been demonstrated to be synchronized by speech stimulus (Luo and Poeppel 2007) and is phase-locked to the speech envelope, i.e., the slow modulations summed over a broad spectral region (Abrams et al. 2008; Ahissar et al. 2001; Aiken and Picton 2008; Lalor and Foxe 2010; Luo and Poeppel 2007). Temporal locking to features of speech has also been supported by intracranial recordings from human core auditory cortex (Nourski et al. 2009). The temporal features of speech contribute significantly to speech intelligibility, as do key spectrotemporal features in speech such as upward and downward formant transitions. The neural coding of spectrotemporal modulations in natural soundtracks has been studied invasively in human auditory cortex using intracranial extracellular recordings (Bitterman et al. 2008), where the spectrotemporal tuning of individual neurons was found to be generally complex and sometimes very fine in frequency. At a neural network level, the blood oxygen level-dependent (BOLD) activity measured by functional magnetic resonance imaging (fMRI) also shows complex spectrotemporal tuning and possesses no obvious spatial map (Schönwiesner and Zatorre 2009). Which spectrotemporal features of speech are encoded in the large-scale synchronized neural activity measurable by MEG and EEG, however, remain unknown and are the focus of the current study.

When investigating the neural coding of speech, there are several key issues that deserve special consideration. One arises from the diversity of speech: language is a productive system permitting the generation of novel sentences. In everyday life, human listeners constantly decode spoken messages they have never heard. In most neurophysiological studies of speech processing, however, small sets of sentences are repeated tens or hundreds of times (although see Lalor and Foxe 2010). This is primarily due to methodological constraints: neurophysiological recordings, especially noninvasive recordings, are quite variable, and so integrating over trials is necessary to obtain a valid estimate of the neural response. An often neglected cost of repeated stimuli, however, is that the listener has obtained complete knowledge of the entire stimulus speech after only a few repetitions. Without the demands of speech comprehension, the encoding of this repeated speech might be quite different from the neural coding of novel speech under natural listening conditions. It is pressing, therefore, to develop experimental paradigms that do not require repeating stimuli many times, to study how speech is encoded in a more ecologically realistic manner.

Second, speech communication is remarkably robust against interference. When competing speech signals are present, human listeners can actively maintain attention on a particular speech target and comprehend it. The superior temporal gyrus has been identified as a region heavily involved in processing

Address for reprint requests and other correspondence: J. Z. Simon, Univ. of Maryland, College Park, MD 20742 (e-mail: jzsimon@umd.edu).

  

concurrent speech signals (Scott et al. 2009). Recent EEG results have shown that human auditory cortex can selectively amplify the low-frequency neural correlates of the speech signal being attended to (Kerlin et al. 2010). This attentional modulation of low-frequency neural activity has been suggested as a general mechanism for sensory information selection (Schroeder and Lakatos 2009). Because speech comprehension is a complex hierarchical process involving multiple brain regions, it is unclear whether the attentional effect seen in the auditory cortex directly modulates feedforward auditory processing or reflects only feedback from language areas, or even motor areas (Hickok and Poeppel 2007). One approach to test whether feedforward processing is involved in speech segregation is to investigate the latency of the attentional effect. If the attentional modulation of MEG/EEG response has a relatively short latency, e.g., 100 ms, then it is evidence that top-down attention modulates representations that are otherwise dominated by feedforward auditory processing. Otherwise, segregating and selectively processing speech may rely on feedback from nonauditory cortex or complex recursive calculations within auditory cortex.

In addition, the auditory encoding of speech is lateralized across the two cerebral hemispheres. It has been hypothesized that the right hemisphere is specialized for the encoding the slow temporal modulations of speech (Poeppel 2003). Support for this hypothesis arises from the observation that neural activity in the right hemisphere is more faithfully synchronized to a speech stimulus than the left, for monaurally and diotically presented speech (Abrams et al. 2008; Luo and Poeppel 2007). Nevertheless, how this proposed intrinsic lateralization of speech encoding interacts with the asymmetry of the ascending auditory pathway is still unclear.

In this study, we investigated the neurophysiology underlying speech processing in human auditory cortex, using minutes-long spoken narratives as stimuli. To address the robustness of this neural coding of speech under more complex listening conditions, the listeners were presented with two simultaneous (and thus competing) spoken narratives, each presented in a separate ear, as a classic, well-controlled illustration of the cocktail party effect (Cherry 1953). This design affords us both the opportunity to investigate the spectrotemporal coding of speech under top-down attentional modulation and the opportunity to separate the intrinsic hemispheric lateralization of speech encoding with the interaction between the left and right auditory pathways. Moreover, previous studies have only demonstrated that speech is encoded in MEG/EEG activity with sufficient fidelity to discriminate among two or three sentences (Kerlin et al. 2010; Luo and Poeppel 2007). With a long-duration, discourse-level stimulus, we can test the limit of this fidelity by quantifying the maximum number of speech stimuli that can be discriminated based on MEG responses.

Inspired by research on single-unit neurophysiology (deCharms et al. 1998; Depireux et al. 2001), the analysis of MEG activity was performed using the spectrotemporal response function (STRF), which can reveal neural coding mechanisms by analyzing the relationship between ongoing neural activity and the corresponding continuous stimuli (Fig. 1). The properties of network-level cortical activity, which plays an important role in auditory processing (Panzeri et al. 2010; Schroeder and Lakatos 2009), were characterized in terms of features of the STRF, such as the spectrotemporal separability (Depireux et al. 2001; Schönwiesner and Zatorre 2009), predictive power (David et al. 2009), binaural composition (Qiu et al. 2003), attentional modulation (Fritz et al. 2003), and hemispheric lateralization, in parallel with what has been done in single-neuron neurophysiology and fMRI. The quantification of these fundamental neurophysiological features establishes the neural strategy used to encode the spectrotemporal features of speech in mass neural activity, conveying information complimentary to that obtained by single-unit neurophysiology and fMRI.

## METHODS

*Subjects.* Ten normal-hearing, right-handed young adults (between 19 and 25 yr old) participated in the experiment, six female. One additional subject participated in the experiment but was excluded from analysis due to excessive head movement (>2 cm) during the experiment. All subjects were paid for their participation. The experimental procedures were approved by the University of Maryland Institutional Review Board. Written informed consent form was obtained from each subject before the experiment.

*Stimuli.* Our stimulus consisted of two segments from a public domain narration of the short story *The Legend of Sleepy Hollow* by Washington Irving (http://librivox.org/the-legend-of-sleepy-hollow-by-washington-irving/), read by a male speaker. The two segments were extracted from different sections of the story, and each of the 2-min-duration segments was further divided into two 1-min-long stimuli. The speech signal was low-pass filtered below 4 kHz. Periods of silence longer than 300 ms were shortened to 300 ms, and white

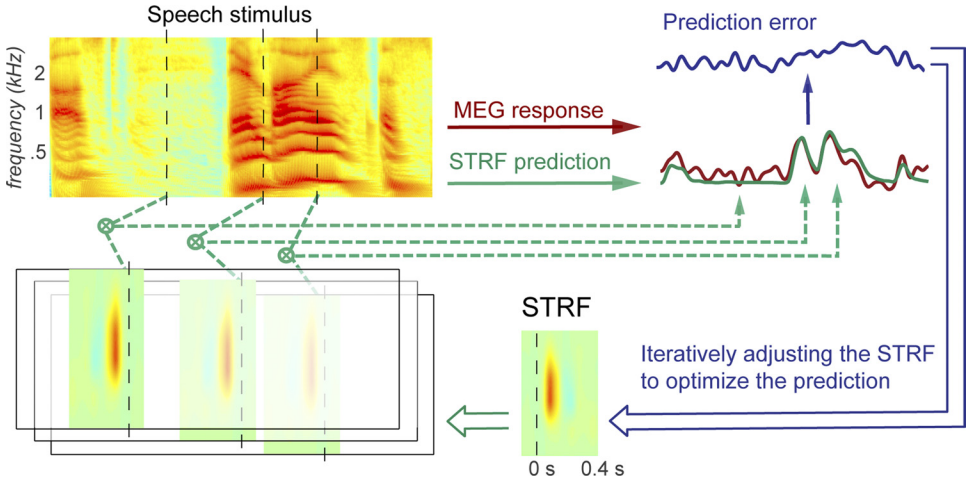


Fig. 1. Illustration of the working principle of the spectrotemporal response function (STRF). The STRF models which stimulus features drive a neural response most powerfully. When an acoustic feature strongly (weakly) resembling the time-reversed STRF appears in the stimulus, the model predicts a strong (weak) neural response. The STRF is iteratively optimized with the use of cross-validation to provide as accurate a prediction of the magnetoencephalography (MEG) response as possible (David et al. 2007).

noise, 20 dB weaker than the speech, was added to the signal to mask any possible subtle discontinuities caused by the removal of silent periods. All stimuli were presented at a comfortable loudness level of around 65 dB. The two stimulus segments were sinusoidally amplitude modulated at 95% modulation depth at 37 and 45 Hz, respectively. As determined by Miller and Licklider (1950), gating a speech signal on and off at a high rate (near 40 Hz) does not significantly affect the intelligibility of speech. Such gating, however, enabled the analysis of auditory steady-state response (aSSR), commonly localized to core auditory cortex (Herdman et al. 2003), and therefore allowed us to monitor the activity in the earliest stage of cortical auditory processing. The association between stimulus segment and modulation rate was counterbalanced over subjects.

*Procedure.* The dichotic listening condition was conducted first. The two audio book excerpts were presented dichotically (separately in each ear) to the subjects using a tube phone plugged into the ear canal. The subjects were instructed to focus on one of the ears until the stimulus ended. The same stimulus was then played again, but the subjects were instructed to switch focus to the other ear. This process was repeated three times for the same set of stimuli, resulting in three identical experimental blocks. All subjects described the dichotic listening task as moderately difficult, and all but one subject reported paying more, or a similar amount of, attention during the second and third presentations of a stimulus, compared with the attention they paid to the first presentation. Which stimulus was played first and which ear was attended to first were counterbalanced over subjects. After the dichotic listening condition was completed, the monaural speech condition was presented. In this condition, each audio book excerpt was presented monaurally, on the same side as in the dichotic condition. Each stimulus was repeated four times. The subjects kept their eyes closed during the whole experiment and had a break every minute. During the break, they were asked a question related to the comprehension of the passage they just heard. On average, the subjects answered 90% of the questions correctly. The performance of the subjects was not significantly different over the three repetitions of the stimulus (1-way repeated-measures ANOVA). In addition, before the main experiment, a pre-experiment was performed. One hundred repetitions of a 500-Hz tone pip were presented to each subject to measure the M100 response.

*Data recording.* The neuromagnetic signal was recorded using a 157-channel whole head MEG system (Kanazawa Institute of Technology, Kanazawa, Japan) in a magnetically shielded room, with a 1-kHz sampling rate. A 200-Hz low-pass filter and a notch filter at 60 Hz were applied online. Three reference channels were used to measure and cancel the environmental magnetic field (de Cheveigne and Simon 2007). Five electromagnetic coils were used to measure each subject's head position inside the MEG machine. The head position was measured twice, once before and once after the experiment, to quantify the head movement during the experiment.

*MEG processing and neural source localization.* Recorded MEG signals contain not only responses directly driven by the stimulus but also stimulus-irrelevant background neural activity. The response component reliably tracking stimulus features is consistent over trials, but the stimulus-irrelevant neural activity is not. On the basis of this property, we decomposed the MEG recording using denoising source separation (DSS) (de Cheveigne and Simon 2008), a blind source separation method that extracts 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. We found that only the first DSS component contains a significant amount of stimulus information (see RESULTS), so analysis was restricted to this component. The spatial magnetic field distribution pattern of this first DSS component was utilized to localize the source of neural responses. In all subjects, the magnetic field corresponding to the first DSS component showed a stereotypical bilateral dipolar pattern and was therefore well modeled by a single equivalent-current dipole (ECD) in each hemisphere. A spherical head model was derived for each subject using the MEG Laboratory software program v.2.001M (Yokogawa Electric, Eagle Technology, Kanazawa Institute of Technology). Position of the ECD was estimated using a global optimization approach (Uutela et al. 1998). The ECD position in each hemisphere was first determined using 54 MEG channels over the corresponding hemisphere. The positions of bilateral ECDs were then refined based on all 157 channels.

After the position of an ECD was determined, the time course of the dipole moment strength was reconstructed using the generalized least-squares method (Mosher et al. 2003). In the reconstructed source activity, the polarity of M100 response was defined as negative (to be consistent with the traditional conventions of MEG/EEG research). The temporal activity reconstructed for the neural sources in the left and right hemispheres was employed for further analysis.

*STRF estimation.* We modeled the cortical auditory processing using the STRF, which describes the input-output relation between a subcortical auditory representation and the cortical MEG response. The subcortical auditory representation of the sounds is a function of frequency and time and is denoted as $S_L(f, t)$ or $S_R(f, t)$ for the stimulus in the left or right ear, respectively. The MEG response is a function of time and is denoted as $r(t)$. The linear STRF model can be formulated as

$$r(t) = \sum_f \sum_\tau \mathrm{STRF}_L(f,\tau) S_L(f, t-\tau) + \sum_f \sum_\tau \mathrm{STRF}_R(f,\tau) S_R(f, t-\tau) + \varepsilon(t),$$

where $\mathrm{STRF}_L(f, t)$ and $\mathrm{STRF}_R(f, t)$ are the STRFs associated with the left- and right-side stimuli and $\varepsilon(t)$ is the residual response waveform not explained by the STRF model. In the monaural stimulus condition, only the relevant stimulus ear is modeled. The subcortical auditory representation is simulated using the model proposed by Yang et al. (1992). This auditory model contains 100 frequency channels between 200 Hz and 4 kHz, similar to a spectrogram in log-frequency scale. For STRF estimation, the 100 frequency channels are downsampled to 5 (David et al. 2007).

The STRF was estimated using boosting with 10-fold cross-validation (David et al. 2007). The estimation procedure is verbally described below, and its strict mathematical formulation is given in the Appendix. During STRF estimation, each 1-min-long MEG response was divided into 10 segments. Nine segments were used to iteratively optimize the STRF (Appendix), whereas the remaining segment was used to evaluate how well the STRF predicts neural responses by its "predictive power": the correlation between MEG measurement and STRF model prediction. Iteration terminated when the predictive power of the STRF decreased for the test segment (e.g., started to demonstrate artifacts of overfitting). The 10-fold cross-validation resulted in 10 estimates of the STRF, whose average was taken as the final result.

*STRF analysis.* The spectral and temporal profiles of the STRF are extracted using singular value decomposition (SVD), $\mathrm{STRF}(f, t) = \Sigma_k \lambda_k \mathrm{SRF}_k(f) \mathrm{TRF}_k(t)$, $\lambda_1 > \lambda_2 > \ldots$. In SVD, the signs of the singular vectors are arbitrary, but we then further require that the spectral singular vectors be overall positive, i.e., $\Sigma_f \mathrm{SRF}_k(f) > 0$. We refer to the first spectral singular vector, i.e., $\mathrm{SRF}_1(f)$, as the normalized spectral sensitivity function and to the product of the first temporal singular vector and its singular value, i.e., $\lambda_1 \mathrm{TRF}_1(t)$, as the temporal response function. The spectral sensitivity function and temporal response function consider only the first spectral and temporal singular vectors, and therefore they only account for some fraction of the total variance of the STRF. This fraction, $\lambda_1^2/\Sigma_k \lambda_k^2$, is called the separability of STRF (Depireux et al. 2001). If the separability of STRF is high (near 1), the STRF is well represented as the outer product of the normalized spectral sensitivity function and the temporal response function, and the spectral and temporal properties of STRF can be discussed separately without any loss of information.

The temporal features of STRF, e.g., the M100-like peak, were extracted from the temporal response function, since the STRF proved to be highly separable. The M100-like peak was determined as the strongest negative peak in the temporal response function between 70 and 250 ms. In analysis of the M100-like response, the MEG responses to each 1-min-long stimulus were averaged within each attentional state unless the experimental block number was employed as an analysis factor.

*Decoding speech information from neural responses.* The STRF model addresses how spectrotemporal features of speech are encoded in cortical neural activity. To test how faithful the neural code was, we employ a decoder to reconstruct the speech features from MEG measurements. Since STRF analyses show that only coarse spectrotemporal modulations of speech are encoded in the MEG activity (see RESULTS), we concentrated on decoding the envelope of speech in a broad frequency band between 400 Hz and 2 kHz (calculated by summing the auditory channels in this range). The linear decoder is formulated as $\hat{s}(t) = \sum_\tau r(t + \tau)D(\tau) + \varepsilon(t)$, where $\hat{s}(t)$, $r(t)$, and $D(t)$ are the decoded speech envelope, the MEG source activity, and the decoder, respectively. This decoding analysis naturally complements the STRF analysis (Mesgarani et al. 2009), and the decoder is estimated using boosting in the same way that the STRF is estimated. The time lag between neural activity and stimulus, $\tau$, is assumed to be between 0 and 500 ms.

To evaluate the performance of the decoder, we calculated the correlation coefficient between the decoded envelope and the envelope of the actual stimulus, and compared it with the correlations between the decoded envelope and the envelopes of other speech signals. We defined the decoding of a neural response as being successfully decoded if the decoded envelope was more correlated with the envelope of the actual stimulus than other nonstimulus envelopes. Using this criterion, when decoding the responses to the four 1-min-duration spoken narratives, a response is correctly decoded if the reconstructed envelope is more correlated with the actual stimulus than the other three stimuli. In this particular case, the decoding task is not very demanding, since only 2 bits of information are needed to discriminate four stimuli while having access to the entire 1-min duration. To test the maximum amount of information decodable from the MEG response, we increased the difficulty of the decoding task by splitting the stimulus and the speech envelope decoded from the neural response into multiple segments and determining the relationship between stimulus and response on a segment basis. For example, if the segment duration is 2 s, each 1-min-long stimulus/response results in 30 segments. To perfectly identify the 30 stimulus segments, one needs at least $\log(30) \simeq 5$ bits of information in the 2-s-long response, resulting in an information rate of 2.5 bit/s (all uses of the log function are with base 2 implied, as is customary in information theoretic analysis). It is worth noting that the information rate in this case describes how faithful the decoded envelope resembles the actual envelope, rather than the linguistic information carried in speech.

Information theory is employed to characterize how much information can be extracted from the neural encoding of speech. The minimal amount of information needed to discriminate $N$ patterns is $\log(N)$ bits. When the mutual information between the stimulus and response, $I(\mathrm{s}, \mathrm{r})$, is less than $\log(N)$ bits, it is not possible to perfectly decode $N$ equally probable stimulus patterns based on the response. The decoding accuracy is limited by Fano's inequality (Cover and Thomas 1991):

$$\mathrm{H}(P_\mathrm{e}) + P_\mathrm{e}\log(N-1) > \log(N) - I(\mathrm{s},\mathrm{r}),$$

where $P_\mathrm{e}$ is percentage of correct decoding and $\mathrm{H}(P_\mathrm{e}) = P_\mathrm{e}\log(P_\mathrm{e}) + (1 - P_\mathrm{e})\log(1 - P_\mathrm{e})$. From the inequality, we also have an estimate of the lower bound of the mutual information between stimulus and response: $I(\mathrm{s}, \mathrm{r}) > \log(N) - \mathrm{H}(P_\mathrm{e}) - P_\mathrm{e}\log(N - 1)$. This inequality holds for any $N$ stimulus patterns, even if the stimulus patterns and the decoding algorithm are optimized. For simplicity, we assume the mutual information $I(\mathrm{s}, \mathrm{r})$ increases linearly with the duration of stimulus/response and therefore express the result as the mutual information rate, mutual information divided by the stimulus duration.

To avoid overfitting while decoding, we divided the two 1-min-long stimuli in each ear into two equal size groups. We used one group to train the decoder and the other group to evaluate decoding accuracy. The two groups were then switched. The decoding results, i.e., the correlation between decoded stimuli and real stimuli, were averaged over the two groups.

*Significance tests.* The statistical significance of the STRF was estimated by comparing the actual STRF results with the null distribution of the STRF parameters. To estimate the null distribution, we derived pseudo-STRFs based on each spoken narrative and mismatched neural responses. To generate a mismatched response, under each listening condition (monaural/attended/unattended), we concatenated all the responses to the four spoken narratives and randomly selected a 1-min-duration neural recording from the concatenated response. One thousand such mismatched responses were generated and resulted in 1,000 pseudo-STRFs in each listening condition.

The predictive power of the actual STRF was viewed as significant if it was greater than any of the predictive powers of the 1,000 pseudo-STRFs ($P < 0.001$). Similarly, the M100-like peak in actual STRF was viewed as significant if it was stronger than any of the peaks in the pseudo-STRF in the same time window ($P < 0.001$). The amplitude of the M100-like response was further analyzed using repeated-measures ANOVA with Greenhouse-Geisser corrections, using the CLEAVE statistical analysis tool (http://www.ebire.org/hcnlab).

*Auditory steady-state response analysis.* Sinusoidal amplitude modulation of a stimulus would be expected to evoke an aSSR at the modulation rate. In the aSSR analysis, responses to the same stimulus were averaged and converted into the frequency domain using the discrete Fourier transform, with 0.017-Hz resolution (based on the 1-min-duration recording). Two stimulus modulation rates, 37 and 45 Hz, were employed in the experiment. In the monaural speech condition, each stimulus was only modulated at one rate, and therefore measurements at the other modulation rate were used to evaluate the background neural noise level at that frequency. The significance of the response at a modulation rate was determined by comparing the response magnitude in the presence of the stimulus modulation and the response magnitude in the absence of the stimulus modulation (permutation test with paired data).

## RESULTS

*Representation of speech in the low-frequency neural response.* In the monaural listening condition, 2 min of a single spoken narrative were presented to each ear. We employed the STRF to model how the spectrotemporal modulations of speech are encoded in the MEG activity filtered into different frequency bands. Figure 2 shows the predictive power of STRF, the correlation between the STRF model prediction and the MEG measurement, for every 2-Hz-wide frequency band between 1 and 59 Hz. The predictive power was above chance level only in the low-frequency region (1–8 Hz), which is further analyzed below.

*Neural representation of spectrotemporal features in speech.* The STRF derived from the low-frequency MEG response (1–8 Hz) is shown in Fig. 3*A*. The STRF can be interpreted in several ways (deCharms et al. 1998; Simon et al. 2007). One is that the STRF at each frequency represents the contribution to the MEG response evoked by a unit power increase of the stimulus in that frequency band. Another, complementary, interpretation is that the STRF, when reversed in time, represents the acoustic features most effective at driving MEG responses (Fig. 1). The STRF shows the strongest activation

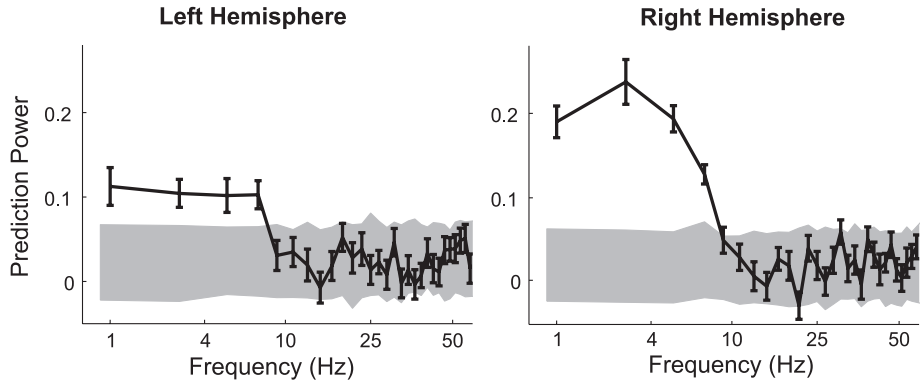


Fig. 2. Predictive power of the STRF model by frequency band. The grand-averaged predictive power is shown as the black line, with error bars representing 1 SE on each side. The gray-shaded area covers from 5 to 95 percentiles of chance level predictive power, estimated based on pseudo-STRFs. The predictive power of STRF of MEG speech response was significantly higher than chance level below 8 Hz.

between 400 Hz and 2 kHz, with a peak at ~100 ms poststimulus. This peak is referred to as the "M100-like response." This STRF indicates that the MEG response tracks spectrotemporal modulations of speech at latency near 100 ms. From another perspective, the instantaneous MEG response is dominantly driven by spectrotemporal modulations that were present in the stimulus 100 ms ago.

The predictive power of STRF was above chance level (test described in METHODS, $P < 0.001$) in each hemisphere for each ear and was significantly higher in the right hemisphere (paired $t$-test, $t_{19} = 3.3$, $P < 0.004$). In the right hemisphere, the grand-averaged predictive power was 0.25 (0.21) for the left (right) side stimulus (significantly higher for the left side, paired $t$-test, $t_9 = 2.4$, $P < 0.05$). In the left hemisphere, the predictive power was similar for stimuli in both ears (0.11 for the left and 0.12 for the right).

An STRF is called spectrotemporally separable when its temporal and spectral processing are independent of each other

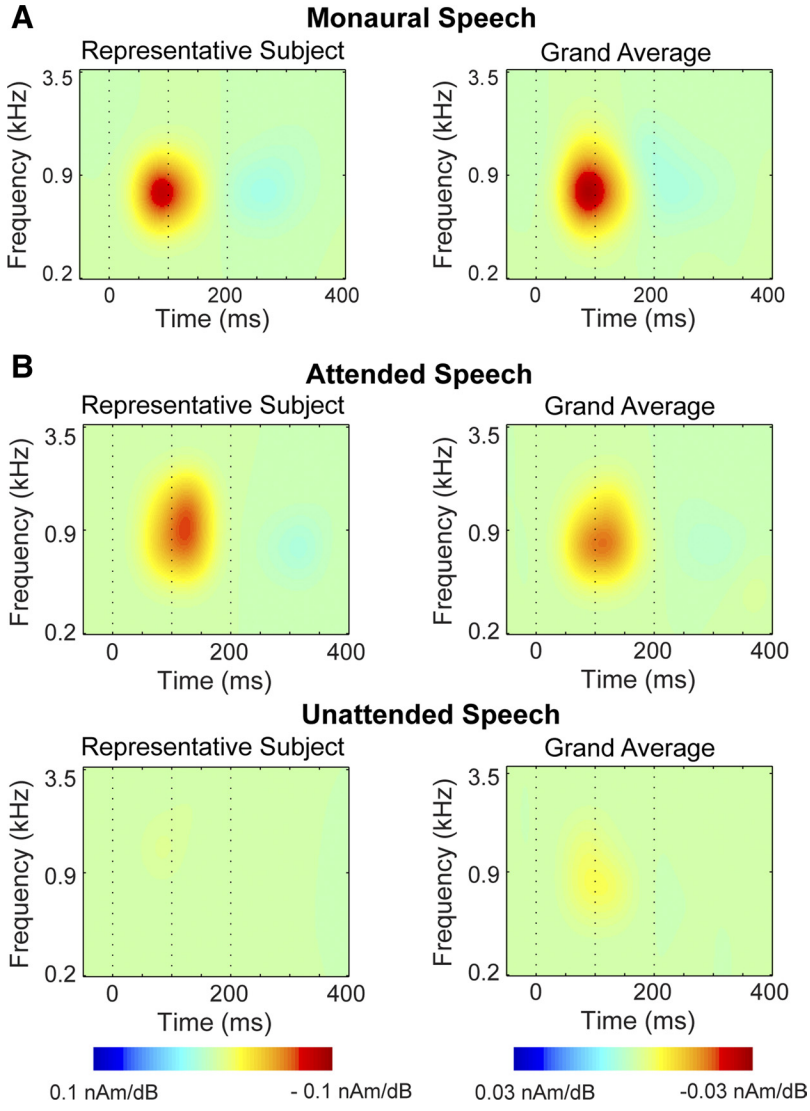


Fig. 3. STRF derived from the MEG speech response to monaurally presented speech (*A*) and dichotically presented simultaneous speech signals (*B*). The most salient feature of the STRF was a negative peak (same polarity as M100/N1) at ~100 ms poststimulus, sometimes followed by a later peak of opposite polarity. In the dichotic listening condition, the amplitude of the STRF was higher for the attended speech than for the interfering (unattended) speech. All examples are from the right hemisphere for speech presented contralaterally. The STRF was smoothed using a 2-dimensional (2-D) Gaussian function with SD of 5 semitones and 25 ms. Data are from representative subject R1474.

(Depireux et al. 2001). The separability of the MEG STRF was very high and is quantitatively illustrated in Fig. 4. Furthermore, the STRF separability was positively correlated with the STRF predictive power (Fig. 4), indicating that STRFs that predict the MEG response well are generally separable. A separable STRF can be decomposed into the product of a single temporal function (Fig. 5*A*) and a single spectral function (Fig. 5*B*), and therefore the spectral property and temporal property of MEG STRFs are analyzed separately as described below.

The normalized spectral sensitivity function of the STRF showed a broad peak between 400 Hz and 2 kHz (Fig. 5*B*). The spectral sensitivity function significantly changed as a function of frequency (frequency × hemisphere × stimulus side, 3-way repeated-measures ANOVA, $F_{1,359} = 28$, $P < 0.0001$) but was not significantly influenced by stimulus side or by hemisphere.

The M100-like response of the STRF was well captured in the temporal response function (Fig. 5*A*) and was statistically significant in each hemisphere for each stimulus presentation side (test described in METHODS, $P < 0.001$). The amplitude and latency of this M100-like response are summarized in Fig. 6. The amplitude of this response was larger in the right hemisphere, independent of the stimulus side (hemisphere × stimulus side, 2-way repeated-measures ANOVA, $F_{1,39} = 11.6$, $P < 0.008$), whereas the latency was shorter for a contralateral stimulus (hemisphere × stimulus side, 2-way repeated-measures ANOVA, $F_{1,39} = 14.6$, $P < 0.005$).

*Speech decoding based on the MEG response.* The STRF analysis above showed that spectrotemporal modulations of speech are encoded in auditory cortex as a temporal code. The fidelity of this temporal code was further assessed by decoding, i.e., reconstructing, speech features from MEG responses. Because the frequency tuning of STRF was broad, we concentrated on decoding the temporal envelope of speech. In the decoding, we divided the MEG response and corresponding stimulus into multiple segments of equal length and used the decoder (estimated from a nonoverlapping data set) to decode the stimulus from each segment. The correlation between the decoded envelope and real stimulus envelope is shown in Fig. 7*A* as a grand-averaged confusion matrix. This result is based on the right hemisphere's response to a 1-min-duration contralateral stimulus for the case where the stimulus and response are divided into 50 (1.2-s duration) segments. In Fig. 7*A*, the 50 stimulus segments and the 50 envelopes decoded from response segments are indexed sequentially from 1 to 50. If each decoded envelope is attributed to the stimulus whose envelope is most correlated with it, 86% of the 50 stimulus segments are correctly decoded.

The number and duration of stimulus/response segments have a profound influence on speech decoding performance. Figure 7*B* shows the speech decoding performance as a function of the number of stimulus segments divided by the duration of each stimulus. Based on Fano's inequality, the speech decoding performance demands that at least 4 bits/s of information in speech were encoded in the right hemisphere MEG response. In the left hemisphere, this value dropped to 1 bit/s. This decoding result is based on the confusion matrix averaged over subjects. Analysis of individual subjects confirms that more information was decoded from the right hemisphere than the left (hemisphere × stimulus side, 2-way

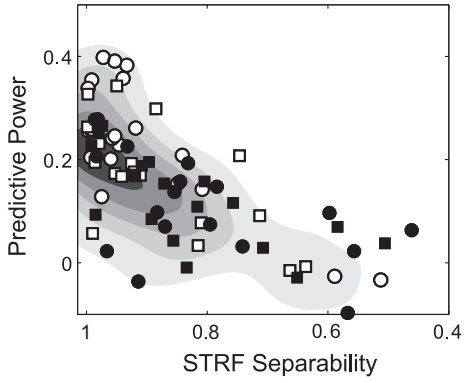


Fig. 4. Predictive power and separability of the STRF. Each point is the result from an individual subject in 1 condition. STRFs with any substantial predictive power are skewed toward high separability. Circles and squares are the results from monaural and binaural listening conditions, respectively; filled and open symbols are results from left and right hemispheres, respectively. The background contour map shows the joint probability distribution density of predictive power and STRF separability. The probability distribution density was obtained by smoothing the 2-D histogram using a Gaussian function (SD = 0.1 in both directions).

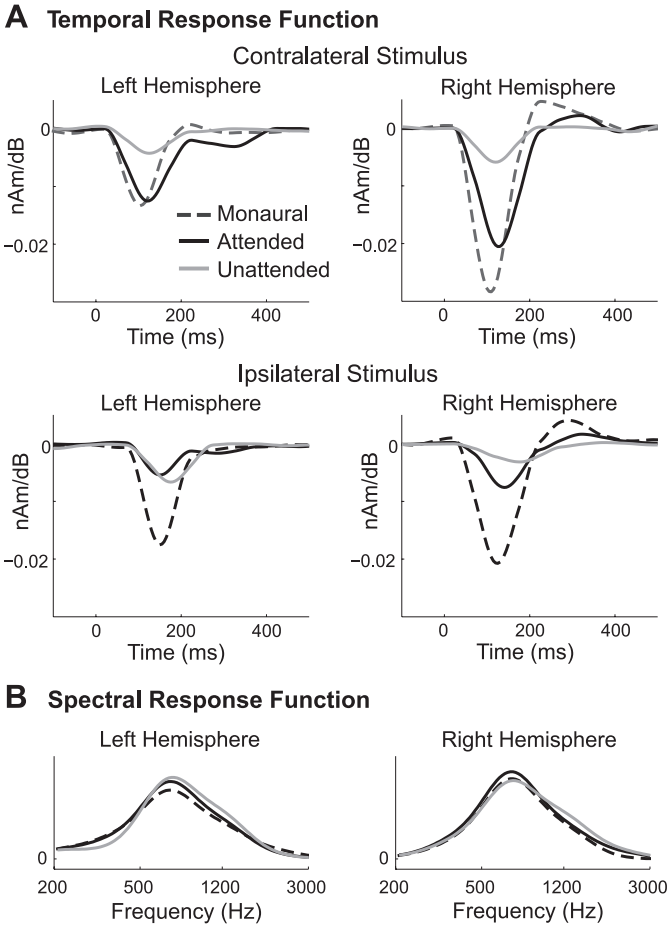


Fig. 5. Temporal response function and spectral sensitivity functions. *A*: grand average of the temporal response functions to speech stimuli under 3 different listening conditions. The amplitude of the temporal response function was higher in the monaural speech condition and was strongly modulated by attention in the dichotic listening condition. *B*: the normalized spectral sensitivity function (grand average over subjects) had a peak between 400 and 2,000 Hz in both hemispheres and all listening conditions. Normalized spectral sensitivity functions to contralateral and ipsilateral stimuli were not significantly different and therefore were averaged. The spectral sensitivity function was smoothed using a Gaussian function with an SD of 5 semitones.

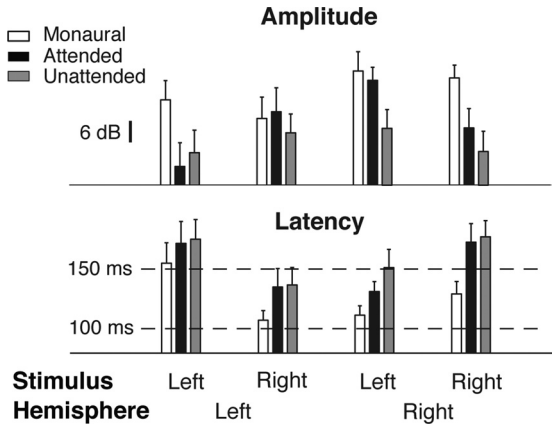


Fig. 6. Amplitude and latency of the M100-like response (grand average). Error bars represent SE. The response amplitude was universally larger and the response latency was universally shorter for monaurally presented speech. In the dichotic condition, the response was stronger for the attended speech than for the unattended speech.

repeated-measures ANOVA, $F_{1,9} = 28.5$, $P < 0.0005$), whereas a similar amount of information was decoded for the left and right side stimuli.

*Spectrotemporal representation of simultaneous speech signals.* Beyond the monaural listening condition analyzed above, subjects also took part in a dichotic listening experiment. In this condition, on top of the single spoken narrative in one ear, another spoken narrative was presented simultaneously in the opposite ear, resulting in a dichotic listening condition. In each experimental block, the subjects were first instructed to listen to the spoken narrative in one ear, and then, when the stimulus was repeated, to listen to the spoken narrative in the other ear. Therefore, the speech signal in each ear served as both a target (when attended to) and an interference signal (when not being attended to). Each experimental block was presented three times. The STRF was determined separately for the stimulus in each ear, under each attentional condition and for each hemisphere.

The STRF for both attended and unattended speech had a salient M100-like response (Figs. 3*B* and 5*A*), similar to the STRF for monaural speech. The STRFs obtained from this dichotic listening condition remained highly separable (Fig. 4). Frequency × hemisphere × attentional state (attended vs. unattended) three-way repeated-measures ANOVA showed that the normalized spectral sensitivity function was not influenced by attentional state and was not different between the two hemispheres (Fig. 5*B*).

The M100-like peak was statistically significant for both attended and unattended speech (test described in METHODS, $P < 0.001$). Compared with the M100-like response for monaural stimuli, the M100-like response to dichotic stimuli was weakened (paired *t*-test, $P \ll 0.0001$ for both attended response and unattended responses) and delayed (paired *t*-test, $P < 0.002$ for attended response and $P \ll 0.0001$ for unattended response). A four-way repeated measures ANOVA (attentional state × hemisphere × stimulus side × experimental block) showed that the latency of this peak in each hemisphere was shorter for the contralateral stimulus ($F_{1,239} = 13.5$, $P < 0.006$).

In the dichotic listening condition, the neural representation of speech remained faithful. The predictive power of the STRF was far above chance level (test described in METHODS, $P < 0.001$). It was not significantly affected by hemisphere or which ear was attended to individually but was affected by the interaction between the two (2-way repeated-measures ANOVA, $F_{1,39} = 20.0$, $P < 0.002$). The predictive power was higher when attention was paid to the contralateral stimulus (0.17 vs. 0.10) for either hemisphere. A considerable amount of speech information can be decoded from the MEG responses to both the attended and the unattended speech. The amount of information extracted from individual subjects was analyzed using a three-way repeated-measures ANOVA (attentional state × hemisphere × stimulus side). More information was decoded when the stimulus was being attended to ($F_{1,79} = 23$, $P < 0.0009$) and in the right hemisphere ($F_{1,79} = 6.5$, $P < 0.03$).

*Attentional modulation during dichotic listening.* The amplitude of this M100-like response peak (Fig. 6) was substantially modulated by attention. A four-way repeated-measures ANOVA (with attentional state, hemisphere, stimulus side, and experimental block number as factors) revealed that the neural response to attended speech was significantly stronger than the neural response to unattended speech ($F_{1,239} = 10.0$, $P < 0.02$). There was a significant interaction among attentional state, hemisphere, and stimulus side ($F_{1,239} = 9.1$, $P < 0.02$). For the speech stimulus in each ear, the attentional effect was more salient in the contralateral hemisphere (paired *t*-test, $t_{59} = 3.3$, $P < 0.002$). There was also an interaction between hemisphere and stimulus side ($F_{1,239} = 16.2$, $P < 0.003$). The response to the stimulus on either side was stronger in the contralateral hemisphere. None of the factors interacted with experimental block number. Even when only the first experimental block was considered, the attention effect was significant (attentional state × hemisphere × stimulus side, 3-way repeated-measures ANOVA, $F_{1,79} = 28.1$, $P < 0.0005$, stronger when attended) and the interaction among attentional state, hemisphere, and stimulus side was significant ($F_{1,79} = 9.0$,

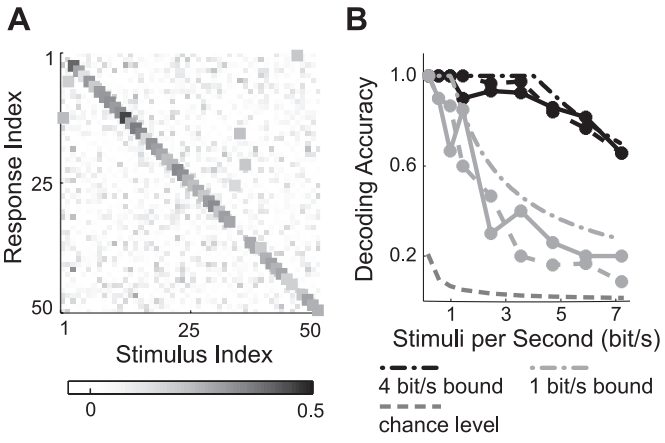


Fig. 7. Stimulus information encoded in the MEG response. *A*: the correlation (grayscale intensity) between the stimulus speech envelope and the envelope reconstructed from the right hemisphere MEG response. The stimulus envelope most correlated with each reconstructed envelope is marked by a square. *B*: stimulus decoding accuracy as a function of the number of stimulus segments per second for monaural speech. The black and gray curves are the results from the left and right hemispheres, respectively; solid and dashed curves are based on the left- and right-side stimuli, respectively. The information decoded from the right and left hemispheres was roughly 4 and 1 bit/s, respectively, for a monaural speech stimulus and is a conservative estimate of the stimulus information available in the MEG response.

$P < 0.02$, attentional modulations stronger in the contralateral hemisphere).

To investigate the temporal dynamics of the attentional gain effect within a single presentation of the stimulus, we divided each 1-min response into ten 6-s segments and estimated the temporal response function for each segment independently. The attentional gain of the M100-like response was extracted from each temporal response function as the gain difference between attended response and unattended response (in dB). A three-way repeated-measures ANOVA (hemisphere × stimulus side × segment) on the attentional gain of the M100-like peak revealed no significant interaction between the attention gain and segment number.

As a result of the attentional gain effect, one might expect the neural response to the speech mixture to be more similar to the neural response to the attended speech than the response to the unattended speech. This hypothesis was confirmed by the analysis of the correlation between the MEG response to the speech mixture and the MEG responses to individual speech components measured during monaural listening (Fig. 8). A three-way repeated-measures ANOVA, with speech component (attended or unattended), hemisphere, and stimulus side as factors, confirmed that the response to the mixture was more correlated with the response to the attended speech component ($F_{1,79} = 36.2$, $P < 0.0002$). The ANOVA analysis also revealed a significant interaction among speech component, hemisphere, and stimulus side ($F_{1,79} = 39.7$, $P < 0.0001$): the response to the mixture was especially dominated by the response to the attended speech in the hemisphere contralateral to the ear the attended speech was presented to.

*Neural source localization.* In the STRF and decoding analyses, the MEG speech response was decomposed into components using a blind source separation method, DSS (de Cheveigne and Simon 2008). Only the first DSS component, which has the strongest trial-to-trial reliability, produced any STRF with substantive predictive power (Fig. 9). The topography of the spatial magnetic field associated with this first DSS component was quantitatively similar to that of the well-known

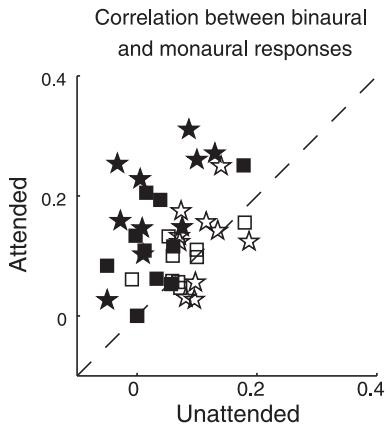


Fig. 8. Correlation between the MEG response to dichotic speech stimuli and the MEG responses to the 2 speech components presented monaurally. Each symbol is the result from 1 subject. The responses in the right and left hemispheres are plotted as stars and squares, respectively. For each hemisphere, if the attended ear in the dichotic condition was the contralateral ear, the result is plotted as a filled symbol, but otherwise it is plotted as an open symbol. The response to dichotic stimuli was more correlated with the response to the attended speech component, especially in the contralateral hemisphere.

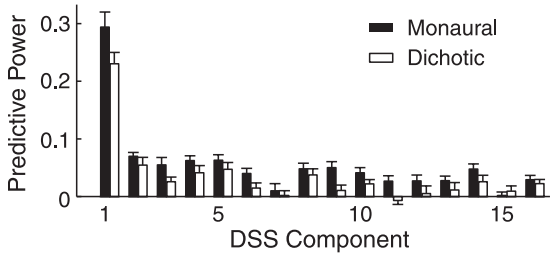


Fig. 9. Predictive power of the STRF derived from each denoising source separation (DSS) component. The first DSS component resulted in significantly higher predictive power than other components and therefore was the only one used to localize the source of the MEG response.

M100 response. The correlation between them was 96.0% for the grand-average magnetic fields (with a 95% confidence interval of 94.6 to 97.0% correlation, estimated by bootstrap sampling). The magnetic field patterns associated with the first DSS component and the M100 were separately modeled by a single ECD in each hemisphere. The correlation between the measured magnetic field and that of the dipole model was 94 ± 5% and 92 ± 7% (mean ± SD) for the DSS component and the M100, respectively. The ECD locations for the two responses were not distinguishable ($P > 0.1$ in all directions), consistent with their topographical similarity, which implies that both are centered in association auditory cortex (Lütkenhöner and Steinsträter 1998; Woldorff et al. 1993).

*Auditory steady-state response.* The sinusoidal modulation of the speech waveforms near 40 Hz generated a small but observable aSSR. For the monaural speech condition, the aSSR at both modulation rates was statistically significant ($P < 0.05$). In the dichotic listening condition, the attentional modulation of aSSR power was assessed by two-way repeated-measures ANOVA (attentional state × hemisphere), but no significant effects were seen.

## DISCUSSION

In this study, we have characterized how spectrotemporal features of speech are encoded in spatially synchronized activity in auditory cortex, by quantifying the relationship between ongoing MEG response and continuous speech stimulus. The major results are summarized as follows. *1*) The neural activity in auditory cortex precisely encodes the slow temporal modulations of speech (<8 Hz) in a broad spectral region between 400 and 2,000 Hz, which roughly encompasses the first and second formants of speech. *2*) The neural coding of slow temporal modulations is stronger and more precise in the right hemisphere, regardless of which ear the speech stimulus is presented to. In the right hemisphere, the neural code is faithful enough to discriminate the responses to hundreds of speech stimuli based on a few seconds of neural recording. *3*) The neural response in each hemisphere is weaker and has a longer latency for speech stimulus monaurally presented to the ipsilateral ear, similar to what is observed for the M100 response (Pantev et al. 1986; Rif et al. 1991).

Using a dichotic listening paradigm, we have further demonstrated how competing speech signals are encoded. *1*) Auditory cortex precisely tracks the temporal modulations of both incoming speech signals but substantially more strongly for the attended one. *2*) The effect of attentional modulation in auditory cortex has a latency of only 100 ms, indicating that the

segregation of dichotic speech stimuli must still involve feedforward neural processing. *3*) The attentional modulation of auditory activity is present even during the subjects' first exposure to a dichotic speech mixture. *4*) The attentional gain effect is more salient in the hemisphere contralateral to the attended ear. *5*) The neural response to speech in either ear is weakened (cf. Fujiki et al. 2002; Penna et al. 2007) and delayed by speech in the other ear.

These results on the spectrotemporal neural encoding of speech provide a clear explanation for stimulus-synchronized neural response observed in previous experiments (Abrams et al. 2008; Ahissar et al. 2001; Aiken and Picton 2008; Lalor and Foxe 2010; Luo and Poeppel 2007; Nourski et al. 2009). The properties and indications of this neural code are discussed below.

*Attentional gain control for unfamiliar speech.* The attention-modulated sensory gain control shown in this study is largely independent of specific knowledge of the content of the speech, since it is effective even on the first exposure to the speech. As far as we know, this is the first evidence that attentional gain modulation is active with a relative short latency when human listeners strive to comprehend novel speech in the presence of interfering speech. Although natural speech has built-in contextual and rhythmic cues, these do not predict the content of the speech by any means. It is known that even without any rhythmic cues, the auditory evoked response to an attended stimulus can be enhanced (Hillyard et al. 1973). It is also possible, however, that contextual cues, and especially the rhythm of natural speech, facilitate the neural amplification of speech encoding (Lakatos et al. 2008). Experiments using dichotically presented tone sequences have demonstrated that the effect of attention on the M100 (N1) is observed for stimuli with some kinds of rhythm (typically fast) (Ahveninen et al. 2011; Hillyard et al. 1973; Power et al. 2011; Rif et al. 1991; Woldorff et al. 1993), but not others (Hari et al. 1989; Ross et al. 2010). Therefore, it is critical to show directly whether early auditory response to speech, with its unique complex temporal structure, is modulated by attention.

Equally as important, the attentional gain effect is seen in auditory cortex, directly affecting a neural response component whose latency is only about 100 ms and which is phase-locked to low-level acoustic features of speech. Therefore, the segregation and selective processing of two dichotically presented speech signals almost certainly involve feedforward auditory neural computations. Also, because of the relatively short latency and the neural source location, it is unlikely that this observed speech segregation occurs during or after the semantic processing of speech. It is also worth noting, however, that the early sensory response to the unattended speech is suppressed but not eliminated. This relatively weak auditory response may be further processed, leading to the interaction between dichotically presented speech signals seen behaviorally (Brungart and Simpson 2002; Conway et al. 2001). In addition, although the M100-like response in STRF is modulated by attention, the aSSR is not. This result is consistent with previous observations that 40-Hz aSSR is not, or only very weakly, modulated by attention or even awareness of sounds (Gutschalk et al. 2008; Lazzouni et al. 2010; Linden et al. 1987). Compared with the M100-like response, the aSSR has a shorter latency at about 50 ms (Ross et al. 2000). Moreover, the neural source location of the aSSR is commonly believed to be in core auditory cortex (Herdman et al. 2003), whereas the neural source location of the M100-like response is centered in association auditory cortex (Lütkenhöner and Steinsträter 1998). Therefore, although feedforward processing is clearly involved in dichotic speech segregation, it may not occur at the level of core auditory cortex. It is also possible, however, that the lack of statistically significant attentional effects on the aSSR is due to the weakness of the aSSR; it is known that aSSR is attenuated by slow temporal modulations, such as those present in speech (Ding and Simon 2009).

Although dichotic speech segregation is reflected in the feedforward early auditory response seen in this study, it is certainly under the modulation of higher order cortical networks. Further experiments are still necessary to identify the network controlling the attentional gain effects seen in auditory cortex, which may include areas in the frontal and parietal cortex (Hill and Miller 2010; Shomstein and Yantis 2006). The attention-control signals by no means need to be phase-locked to acoustic features of the speech stimulus and therefore cannot be extracted using the STRF analysis employed presently.

In addition, since the current experiment uses the same speaker and same narrative source for both ears, rather than tones of different frequencies, we have demonstrated that this attentional sensory gain control can be driven entirely by the stimulus ear, not needing, for example, spectral cues. Of course, other monaural cues, such as pitch and rhythm, and binaural cues, such as interaural time difference (ITD) and interaural level difference (ILD), can also be utilized to segregate concurrent sounds (Bronkhorst 2000). Previous experiments with simple nonspeech stimuli have demonstrated that monaural cue-based segregation of spectrally nonoverlapping sounds is reflected neurally in human auditory cortex (Bidet-Caulet et al. 2007; Elhilali et al. 2009; Xiang et al. 2010). Future experiments are needed to address whether speech segregation itself, which is a much harder problem, also occurs in human auditory cortex at a short latency.

*Hemispheric lateralization of speech coding in auditory cortex.* Although the neural tracking of spectrotemporal modulations in speech is seen bilaterally, it is strongly lateralized to the right hemisphere, independent of the stimulus ear. This lateralization is demonstrated by the amplitude of the M100-like component (Fig. 6) and more critically by the fidelity of neural coding (Fig. 7B). This strong right hemisphere dominance effect is surprising, however, since it is not observed in the M100 response to sound onsets or to aSSR to 40-Hz amplitude modulations (Rif et al. 1991; Ross et al. 2005), both of which are instead stronger in the hemisphere contralateral to the ear receiving the stimulus or equally strong in both hemispheres. Furthermore, even for responses to speech, if both the response tracking speech features and other responses are considered, the total response is stronger in the left rather than right hemisphere (Millman et al. 2011). Nor can the rightward lateralization of the neural representation of speech be explained anatomically, since the dominant excitatory input to each hemisphere is from the contralateral ear (Pickles 1988). Therefore, this result gives further support to the hypothesis that the right hemisphere is intrinsically dominant in processing the slow modulations ($<$10 Hz) in speech during natural speech comprehension (Poeppel 2003). This right hemisphere dominance has also been observed in the neural response to speech (Abrams et al. 2008; Kerlin et al. 2010; Luo and

Poeppel 2007) and even in endogenous neural oscillations (Giraud et al. 2007).

On top of this intrinsic right hemisphere dominance, however, during dichotic listening, the effect of attentional gain control is even more prominent in the hemisphere contralateral to the attended side. This hemispheric lateralization effect likely arises from the anatomical asymmetry between the left and right afferent pathways to each hemisphere. When two different sounds are presented to the two ears separately, their neural representations form a competition (Fujiki et al. 2002; Penna et al. 2007). One result of this competition may be that each hemisphere primarily processes information from the contralateral ear, where most of the excitatory afferent inputs are from (Pickles 1988). Therefore, the neural processing of each stimulus can be most strongly modulated by the attentional gain change in the contralateral hemisphere.

*Neural coding of spectrotemporal dynamics of speech signals.* Using STRF analysis, we have demonstrated that slow temporal modulations of speech (particularly of coarse spectral modulations) are precisely encoded in human auditory cortex. Taking advantage of the fine time resolution of MEG, we have shown that the observed neural responses encode at least 4 bit/s information. This indicates that, using a linear decoder, we can errorlessly discriminate about 16 speech stimuli (4 bits) of 1-s duration based on their MEG responses. Similarly, this same information rate allows one to errorlessly discriminate about 256 speech stimuli (8 bits) of 2-s duration. The possibility of discriminating MEG/EEG responses to speech has been suggested by earlier studies but only shown on the basis of a small number of sentences of several seconds in duration (Kerlin et al. 2010; Luo and Poeppel 2007). The MEG response is also robust: an M100-like response is observed even for unattended speech. This contrasts with the observation that the neural representation of sounds in anesthetized avian auditory forebrain is severely degraded by acoustic interference (Narayan et al. 2007) and therefore suggests that the robust neural coding may require top-down attentional modulation.

In speech, temporal modulations below 10 Hz convey syllabic and phrasal level information (Greenberg 1999). In quiet, these slow modulations, in concert with even a very coarse spectral modulation, accomplish high speech intelligibility (Shannon et al. 1995). When speech is masked by acoustic interference, slow temporal modulations of the interference releases the masking of the target speech (Festen and Plomp 1990). Faster acoustic fluctuations of speech, e.g., spectral and pitch cues, that contain phonetic and prosodic information are gated by the slow modulations (Rosen 1992). Similarly, the neural processing of speech features on short time scales ($<$100 ms) may also be modulated by the low-frequency neural activity analyzed in this study. The phonetic information of speech has been suggested to be spatially coded over neural populations in auditory cortex (Chang et al. 2010). This spatial code discriminates different syllables most effectively at around 100 ms after the syllable onset, consistent with the latency of the M100-like response in the MEG STRF. Other possible neural signatures of higher level processing of speech are high-frequency neural oscillations (40–150 Hz), which are also coupled to slow neural oscillations below 10 Hz (Lakatos et al. 2008). Therefore, the slow activity noninvasively measured by MEG probably reflects the timing of such microscopic neural computations of the phonetic level information of speech.

*The STRF of the MEG speech response.* The mathematical linear system bridging the speech stimulus and the neural representation of that speech can be represented graphically by the STRF. The predictive power of the MEG STRF compares well with that obtained from single cortical neurons for speech stimuli (Bitterman et al. 2008; David et al. 2007, 2009). The MEG STRF is highly separable: the temporal processing of the speech stimulus is consistent over the entire frequency range of the STRF and is equally sensitive to upward and downward changes in frequency content. This contrasts with the variety of separability seen in the STRFs of single neurons in primary auditory cortex (Depireux et al. 2001) and the inseparability seen using fMRI (Schönwiesner and Zatorre 2009). This difference in separability reflects differences between the spectrotemporal tuning of individual neurons, spatially synchronized activity and nonspatially synchronized activity. MEG and fMRI recordings reflect the activity of large neural populations. In addition, MEG records only spatially synchronized components of the response (and in this study, stimulus-synchronized neural activity), whereas fMRI measures the indirect hemodynamic response, which is influenced by both synchronized and asynchronous neural activity. Hence, MEG and fMRI demonstrate very different aspects of the population level distribution of the spectrotemporal tuning properties of neurons and are therefore naturally complementary.

In summary, in this study we have demonstrated the existence of a neural encoding of speech in human auditory cortex that can be measured extracranially and noninvasively. We also have demonstrated that this neural encoding is based on the acoustic modulations of the spectrotemporal features of speech. The encoding is quite faithful (perhaps even surprisingly so given that the neural signal is measured extracranially) and is able to distinguish among hundreds of different stimuli in the course of only a few seconds. In addition, on one hand, the encoding strategy is very strongly tied to the physical properties of speech, which would normally imply a bottom-up encoding process, but on the other hand, the encoding strategy is also strongly modulated by the attentional state of the listener, demonstrating that top-down processes directly modulate the neural representation of the fundamental acoustic features of speech. Finally, we also have developed a practical experimental paradigm that allows single-trial analysis of the auditory cortical encoding of continuous speech in an ecologically realistic manner.

## APPENDIX

The STRF based on the MEG response is estimated using boosting (David et al. 2007) with 10-fold cross-validation. The procedure is documented as follows:

*1*) Initialize the STRF.

$$\mathrm{STRF}_0(f, t) = 0, \text{ for all } f \text{ and } t.$$

*2*) Iteratively optimize the STRF. The $n$th iteration is based on the results of the $(n - 1)$th iteration:

$$r_{n-1}(t) = \sum_f \sum_\tau \mathrm{STRF}_{n-1}(f, \tau) S(f, t - \tau) + \varepsilon_{n-1}(t).$$

In the $n$th iteration,

$$r_n(t) = \sum_f \sum_\tau \mathrm{STRF}_n(f,\tau) S(f, t-\tau) + \varepsilon_n(t),$$

where

$$\mathrm{STRF}_n(f,\tau) = \mathrm{STRF}_{n-1}(f,\tau) + \Delta\mathrm{STRF}(f,\tau),$$

$$\Delta\mathrm{STRF}(f,\tau) = \begin{cases} \delta, \text{ if } f = f_0, t = t_0 \\ 0 \end{cases}.$$

The prediction error in the $n$th iteration is $\varepsilon_n(t) = \varepsilon_{n-1}(t) - \delta S(f_0, t_0)$. $\Delta$STRF is selected to minimize the prediction error, i.e.,

$$\Delta\mathrm{STRF}(f,\tau) = \underset{f_0,t_0}{\operatorname{argmin}} \sum_t \varepsilon_n^2(t) = \underset{f_0,t_0}{\operatorname{argmin}} \sum_t \left[\varepsilon_{n-1}(t) - \delta S(f_0,t_0)\right]^2.$$

*3*) Terminate the iteration when the prediction error of the model drops based on cross-validation.


## ACKNOWLEDGMENTS

We thank D. Poeppel and M. F. Howard for valuable feedback on earlier versions of this manuscript, S. V. David and S. A. Shamma for insightful discussions on data analysis, and M. Ehrmann for excellent technical assistance.

## GRANTS

This research was supported by the National Institute of Deafness and Other Communication Disorders Grants R01 DC-008342 and R01 DC-005660.


## DISCLOSURES

No conflicts of interest, financial or otherwise, are declared by the author(s).

## AUTHOR CONTRIBUTIONS

N.D. and J.Z.S. conception and design of research; N.D. performed experiments; N.D. and J.Z.S. analyzed data; N.D. and J.Z.S. interpreted results of experiments; N.D. and J.Z.S. prepared figures; N.D. and J.Z.S. drafted manuscript; N.D. and J.Z.S. edited and revised manuscript; N.D. and J.Z.S. approved final version of manuscript.

## REFERENCES


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

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

**Ahveninen J, Hämäläinen M, Jääskeläinen IP, Ahlfors SP, Huang S, Lin FH, Raij T, Sams M, Vasios CE, Belliveau JW.** Attention-driven auditory cortex short-term plasticity helps segregate relevant sounds from noise. *Proc Natl Acad Sci USA* 108: 4182–4187, 2011.

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

**Bidet-Caulet A, Fischer C, Besle J, Aguera PE, Giard MH, Bertrand O.** Effects of selective attention on the electrophysiological representation of concurrent sounds in the human auditory cortex. *J Neurosci* 27: 9252–9261, 2007.

**Bitterman Y, Mukamel R, Malach R, Fried I, Nelken I.** Ultra-fine frequency tuning revealed in single neurons of human auditory cortex. *Nature* 451: 197–201, 2008.

**Bronkhorst AW.** The cocktail party phenomenon: a review of research on speech intelligibility in multiple-talker conditions. *Acustica* 86: 117–128, 2000.

**Brungart DS, Simpson BD.** Within-ear and across-ear interference in a cocktail-party listening task. *J Acoust Soc Am* 122: 2985–2995 2002.

**Chang E, Rieger J, Johnson K, Berger M, Barbaro N, Knight R.** Categorical speech representation in human superior temporal gyrus. *Nat Neurosci* 13: 1428–1432, 2010.

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

**Conway ARA, Cowan N, Bunting MF.** The cocktail party phenomenon revisited: the importance of working memory capacity. *Psychon Bull Rev* 8: 331–335, 2001.

**Cover TM, Thomas JA.** *Elements of Information Theory*. New York: Wiley, 1991.

**David SV, Mesgarani N, Fritz JB, Shamma SA.** Rapid synaptic depression explains nonlinear modulation of spectro-temporal tuning in primary auditory cortex by natural stimuli. *J Neurosci* 29: 3374–3386, 2009.

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

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

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

**deCharms RC, Blake DT, Merzenich MM.** Optimizing sound features for cortical neurons. *Science* 280: 1439–1444, 1998.

**Depireux DA, Simon JZ, Klein DJ, Shamma SA.** Spectro-temporal response field characterization with dynamic ripples in ferret primary auditory cortex. *J Neurophysiol* 85: 1220–1234, 2001.

**Ding N, Simon JZ.** Neural representations of complex temporal modulations in the human auditory cortex. *J Neurophysiol* 102: 2731–2743, 2009.

**Elhilali M, Xiang J, Shamma SA, Simon JZ.** Interaction between attention and bottom-up saliency mediates the representation of foreground and background in an auditory scene. *PLoS Biol* 7: e1000129, 2009.

**Festen JM, Plomp R.** Effects of fluctuating noise and interfering speech on the speech-reception threshold for impaired and normal hearing. *J Acoust Soc Am* 88: 1725–1736, 1990.

**Fritz J, Shamma S, Elhilali M, Klein D.** Rapid task-related plasticity of spectrotemporal receptive fields in primary auditory cortex. *Nat Neurosci* 6: 1216–1223, 2003.

**Fujiki N, Jousmäki V, Hari R.** Neuromagnetic responses to frequency-tagged sounds: a new method to follow inputs from each ear to the human auditory cortex during binaural hearing. *J Neurosci* 22: RC205, 2002.

**Giraud AL, Kleinschmidt A, Poeppel D, Lund TE, Frackowiak RSJ, Laufs H.** Endogenous cortical rhythms determine cerebral specialization for speech perception and production. *Neuron* 56: 1127–1134, 2007.

**Greenberg S.** Speaking in shorthand—a syllable-centric perspective for understanding pronunciation variation. *Speech Commun* 29: 159–176, 1999.

**Gutschalk A, Micheyl C, Oxenham AJ.** Neural correlates of auditory perceptual awareness under informational masking. *PLoS Biol* 6: e138, 2008.

**Hari R, Hämäläinen M, Kaukoranta E, Mäkelä J, Joutsiniemi SL, Tiihonen J.** Selective listening modifies activity of the human auditory cortex. *Exp Brain Res* 74: 463–470, 1989.

**Herdman AT, Wollbrink A, Chau W, Ishii R, Ross B, Pantev C.** Determination of activation areas in the human auditory cortex by means of synthetic aperture magnetometry. *Neuroimage* 20: 995–1005, 2003.

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

**Hill KT, Miller LM.** Auditory attentional control and selection during cocktail party listening. *Cereb Cortex* 20: 583–590, 2010.

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

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

**Lakatos P, Karmos G, Mehta AD, Ulbert I, Schroeder CE.** In monkeys that are paying attention to a rhythmic stimulus, brain oscillations become tuned to the stimulus so that the response in the visual cortex is enhanced. *Science* 320: 110–113, 2008.

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

**Lazzouni L, Ross B, Voss P, Lepore F.** Neuromagnetic auditory steady-state responses to amplitude modulated sounds following dichotic or monaural presentation. *Clin Neurophysiol* 121: 200–207, 2010.

**Linden RD, Picton TW, Hamel G, Campbell KB.** Human auditory steady-state evoked potentials during selective attention. *Electroencephalogr Clin Neurophysiol* 66: 145–159, 1987.

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

**Lütkenhöner B, Steinsträter O.** High-precision neuromagnetic study of the functional organization of the human auditory cortex. *Audiol Neurootol* 3: 191–213, 1998.

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

**Miller GA, Licklider JC.** The intelligibility of interrupted speech. *J Acoust Soc Am* 22: 167–173, 1950.

**Millman RE, Woods WP, Quinlan PT.** Functional asymmetries in the representation of noise-vocoded speech. *Neuroimage* 54: 2364–2373, 2011.

**Mosher JC, Baillet S, Leahy RM.** Equivalence of linear approaches in bioelectromagnetic inverse solutions. In: *IEEE Workshop on Statistical Signal Processing*. St. Louis, MO: 2003.

**Narayan R, Best V, Ozmeral E, McClaine E, Dent M, Shinn-Cunningham B, Sen K.** Cortical interference effects in the cocktail party problem. *Nat Neurosci* 10: 1601–1607, 2007.

**Nourski KV, Reale RA, Oya H, Kawasaki H, Kovach CK, Chen H, Matthew A, Howard I, Brugge JF.** Temporal envelope of time-compressed speech represented in the human auditory cortex. *J Neurosci* 29: 15564–15574, 2009.

**Pantev C, Lütkenhöner B, Hoke M, Lehnertz K.** Comparison between simultaneously recorded auditory-evoked magnetic fields and potentials elicited by ipsilateral, contralateral and binaural tone burst stimulation. *Audiology* 25: 54–61, 1986.

**Panzeri S, Brunel N, Logothetis NK, Kayser C.** Sensory neural codes using multiplexed temporal scales. *Trends Neurosci* 33: 111–120, 2010.

**Penna SD, Brancucci A, Babiloni C, Franciotti R, Pizzella V, Rossi D, Torquati K, Rossini PM, Romani GL.** Lateralization of dichotic speech stimuli is based on specific auditory pathway interactions: neuromagnetic evidence. *Cereb Cortex* 17: 2303–2311, 2007.

**Pickles JO.** *An Introduction to the Physiology of Hearing*. New York: Academic, 1988.

**Poeppel D.** The analysis of speech in different temporal integration windows: cerebral lateralization as 'asymmetric sampling in time'. *Speech Commun* 41: 245–255, 2003.

**Power AJ, Lalor EC, Reilly RB.** Endogenous auditory spatial attention modulates obligatory sensory activity in auditory cortex. *Cereb Cortex* 21: 1223–1230, 2011.

**Qiu A, Schreiner CE, Escabí MA.** Gabor Analysis of Auditory Midbrain Receptive Fields: Spectro-Temporal and Binaural Composition. *J Neurophysiol* 90: 456–476, 2003.

**Rif J, Hari R, Hämäläinen MS, Sams M.** Auditory attention affects two different areas in the human supratemporal cortex. *Electroencephalogr Clin Neurophysiol* 79: 464–472, 1991.

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

**Ross B, Borgmann C, Draganova R, Roberts LE, Pantev C.** A high-precision magnetoencephalographic study of human auditory steady-state responses to amplitude-modulated tones. *J Acoust Soc Am* 108: 679–691, 2000.

**Ross B, Herdman AT, Pantev C.** Right hemispheric laterality of human 40 Hz auditory steady-state responses. *Cereb Cortex* 15: 2029–2039, 2005.

**Ross B, Hillyard SA, Picton TW.** Temporal dynamics of selective attention during dichotic listening. *Cereb Cortex* 20: 1360–1371, 2010.

**Schönwiesner M, Zatorre RJ.** Spectro-temporal modulation transfer function of single voxels in the human auditory cortex measured with high-resolution fMRI. *Proc Natl Acad Sci USA* 106: 14611–14616, 2009.

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

**Scott SK, Rosen S, Beaman CP, Davis JP, Wise RJ.** The neural processing of masked speech: evidence for different mechanisms in the left and right temporal lobes. *J Acoust Soc Am* 125: 1737–1743, 2009.

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

**Shomstein S, Yantis S.** Parietal cortex mediates voluntary control of spatial and nonspatial auditory attention. *J Neurosci* 26: 435–439, 2006.

**Simon JZ, Depireux DA, Klein DJ, Fritz JB, Shamma SA.** Temporal symmetry in primary auditory cortex: implications for cortical connectivity. *Neural Comput* 19: 583–638, 2007.

**Uutela K, Hämäläinen M, Salmelin R.** Global optimization in the localization of neuromagnetic sources. *IEEE Trans Biomed Eng* 45: 716–723, 1998.

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

**Xiang J, Simon J, Elhilali M.** Competing streams at the cocktail party: exploring the mechanisms of attention and temporal integration. *J Neurosci* 30: 12084–12093, 2010.

**Yang X, Wang K, Shamma SA.** Auditory representations of acoustic signals. *IEEE Trans Info Theory* 38: 824–839, 1992.