Suppression of competing speech through entrainment of cortical oscillations

Cort Horton, Michael D'Zmura and Ramesh Srinivasan

J Neurophysiol 109:3082-3093, 2013. First published 20 March 2013;
doi: 10.1152/jn.01026.2012

J Neurophysiol 109: 3082–3093, 2013.
First published March 20, 2013; doi:10.1152/jn.01026.2012.

Suppression of competing speech through entrainment of cortical oscillations

Cort Horton,[1] Michael D'Zmura,[1] and Ramesh Srinivasan[1,2]
[1]*Department of Cognitive Sciences, University of California, Irvine, California; and* [2]*Department of Biomedical Engineering, University of California, Irvine, California*

Submitted 27 November 2012; accepted in final form 13 March 2013

Horton C, D'Zmura M, Srinivasan R. Suppression of competing speech through entrainment of cortical oscillations. *J Neurophysiol* 109: 3082–3093, 2013. First published March 20, 2013; doi:10.1152/jn.01026.2012.—People are highly skilled at attending to one speaker in the presence of competitors, but the neural mechanisms supporting this remain unclear. Recent studies have argued that the auditory system enhances the gain of a speech stream relative to competitors by entraining (or "phase-locking") to the rhythmic structure in its acoustic envelope, thus ensuring that syllables arrive during periods of high neuronal excitability. We hypothesized that such a mechanism could also suppress a competing speech stream by ensuring that syllables arrive during periods of low neuronal excitability. To test this, we analyzed high-density EEG recorded from human adults while they attended to one of two competing, naturalistic speech streams. By calculating the cross-correlation between the EEG channels and the speech envelopes, we found evidence of entrainment to the attended speech's acoustic envelope as well as weaker yet significant entrainment to the unattended speech's envelope. An independent component analysis (ICA) decomposition of the data revealed sources in the posterior temporal cortices that displayed robust correlations to both the attended and unattended envelopes. Critically, in these components the signs of the correlations when attended were opposite those when unattended, consistent with the hypothesized entrainment-based suppressive mechanism.

EEG; independent component analysis; selective attention; speech envelopes; phase-locking

PEOPLE ARE FREQUENTLY TASKED with selectively attending to a speaker in the presence of competing speakers, commonly described as the "cocktail party problem" (Cherry 1953). Listeners are remarkably skilled at this—utilizing spatial, spectral, temporal, semantic, and even visual cues (when available) to segregate the desired speech from the complex mixture (Bronkhorst 2000). Considering the ease with which people do this, it has been surprisingly difficult for researchers to describe the neural mechanisms that support this ability (McDermott 2009).

While the mechanisms of selective attention to a speaker may be uncertain, the effects on neural activity can be readily measured with functional imaging techniques. Researchers using electroencephalography (EEG) and magnetoencephalography (MEG) have long found that simple auditory stimuli elicit larger evoked responses when attended than when unattended (Picton and Hillyard 1974; Woldorff et al. 1993). Recent studies using continuous measures, more appropriate for assessing neural responses to ongoing naturalistic speech stimuli, have followed essentially the same pattern (Ding and Simon 2012a; Kerlin et al. 2010; Mesgarani and Chang 2012; Power et al. 2011, 2012). Selectively attending to a speaker appears to increase the gain of the attended speech and/or reduce the gain of the competing speech. However, the manner in which this attentional gain is applied may depend on the cues used to distinguish between speakers. For example, if competing speakers differ in the spectral content of their voices, gain could be applied directly to corresponding tonotopic populations (Alcaini et al. 1995; Da Costa et al. 2013). Likewise, if speakers are in different locations in space, gain could be applied to sounds with specific interaural timing and level differences (Darwin and Hukin 1999; Nager et al. 2003).

For temporal cues, a mechanism for applying attentional gain has recently been proposed that relies on the hierarchical structure of oscillations in sensory cortex (Lakatos et al. 2008; Schroeder and Lakatos 2009). The amplitudes of gamma-band (30+ Hz) oscillations, which are strongly linked to spiking activity and local processing, are dependent on the phases of delta (<4 Hz)- and theta (4–8 Hz)-band oscillations, which affect cyclical changes in population excitability (Canolty et al. 2006; Lakatos et al. 2005, 2008; Schroeder and Lakatos 2009). Stimuli that arrive during excited phases evoke more spiking activity, produce stronger bursts of gamma oscillations, and have faster reaction times than stimuli that arrive in less excited phases (Lakatos et al. 2008). In and of itself, this hierarchy does not promote attentional selection. However, with sufficiently rhythmic input, the endogenous oscillatory activity in sensory cortex can entrain to the stimuli (Walter and Walter 1949; Will and Berg 2007), so that each successive stimulus falls into this high-excitability phase. Lakatos and colleagues argued that this entrainment provides a means for sensory selection; when multiple rhythmic stimuli are in competition, the gain of the attended stimulus can be enhanced by encouraging entrainment to its rhythm instead of those of competitors (Lakatos et al. 2008; Schroeder and Lakatos 2009).

This temporal attention mechanism is well-suited for use in "cocktail party" scenarios. Speech contains rhythmic structure at both of the timescales that Lakatos and colleagues identified as controlling neural population excitability; prosody and phrasing impart slow (delta band) modulations in the intensity of speech utterances, while the boundaries of syllables are encoded in theta-band modulations (Poeppel 2003). Collectively, these low-frequency modulations in the intensity of speech utterances over time are referred to as the acoustic (or temporal) envelope of the speech (Rosen 1992). By entraining (or "phase-locking") to the theta-band modulations in the envelope, populations in auditory cortex would be particularly excitable at the times when exogenous activity from new syllables reached their input layers. Similarly, entraining to the delta-band modulations in the envelope could raise neural

Address for reprint requests and other correspondence: C. Horton, Dept. of Cognitive Sciences, 2201 Social & Behavioral Sciences Gateway Bldg. (SBSG), Univ. of California, Irvine, CA 92697-5100 (e-mail: chorton@uci.edu).

excitability at stressed periods within sentences. In both cases, the increased excitability should give rise to larger neural responses to the attended speech, and thus a better representation within cortical speech processing areas. Even in the absence of competitors, some researchers have argued that entrainment to these modulations is a fundamental feature of cortical speech processing (Ghitza 2011; Giraud and Poeppel 2012).

We hypothesized that this temporal attention mechanism should also be capable of suppressing a competing speaker by keeping auditory activity 180° out of phase with the competing speech's acoustic envelope, thus ensuring that syllables would arrive during periods of reduced neuronal excitability. The role of suppression as a mechanism underlying selective attention is well-established in vision (Desimone and Duncan 1995; Hopf et al. 2006), where distinct parietal networks have been shown to mediate enhancement and suppression (Bridwell and Srinivasan 2012). With auditory stimuli, the role of suppression in selective attention is somewhat less clear. Although both enhancement and suppression of neural responses have been observed during the engagement of auditory selective attention (Hillyard et al. 1998), increased evoked response magnitudes have also been interpreted as evidence of "active" suppression (Rif et al. 1991). Here, we only refer to suppression in terms of a reduction in evoked responses.

Data from functional imaging studies support the idea of enhancement of speech through entrainment. Oscillations in auditory cortex have been shown to phase-lock to the envelope of speech when no competing streams are present (Abrams et al. 2008; Ahissar et al. 2001; Aiken and Picton 2008; Howard and Poeppel 2010; Lalor and Foxe 2010; Luo and Poeppel 2007; Nourski et al. 2009; Pasley et al. 2012). When competing streams are introduced, neural activity tracks the envelope of the attended speech better than that of the unattended speech—comparably to when there were no competitors (Ding and Simon 2012a, 2012b; Kerlin et al. 2010; Mesgarani and Chang 2012). That pattern is consistent with an enhancement of the attended speech, a suppression of the unattended speech, or both. It is not consistent, however, with an entrainment-based suppressive mechanism. If the hypothesized suppressive mechanism had been utilized, auditory activity would have remained opposite in phase to the competing speech, encoding the inverse of the competing stream's envelope. Currently, we are unaware of any published results that fit that pattern, but the behavioral tasks used to date have not been specifically designed to invoke suppression.

In the present study, we attempted to create a situation in which it was imperative that the unattended speech be suppressed so that we might see evidence of suppression through entrainment. We recorded high-density EEG while human subjects attended to one of two competing speakers during a naturalistic "cocktail party" task. To encourage the suppression of the nontarget speaker, we implemented a behavioral task that required subjects to completely "tune out" the competing speech in order to succeed. We then used cross-correlation to measure phase-locking between the EEG and the acoustic envelopes of the attended and unattended speech. Additionally, we used independent component analysis (ICA) to distinguish the activity of individual sources within the brain, as speech processing takes place within a relatively large network of cortical areas (Hickok and Poeppel 2007)—any of which could potentially phase-lock to the speech envelopes. Finally, we amplitude-modulated the speech stimuli at high frequencies in order to induce auditory steady-state responses (ASSRs), which have been used to measure changes in auditory stimulus gain resulting from attention (Bidet-Caulet et al. 2007; Lazzouni et al. 2010; Linden et al. 1987; Müller et al. 2009; Ross et al. 2004; Skosnik et al. 2007). Our findings suggest that neural entrainment is used in both enhancement and suppression of speech in "cocktail party" scenarios.

MATERIALS AND METHODS

This study was conducted according to protocols reviewed and approved by the Institutional Review Board of the University of California, Irvine, and all participants gave written informed consent prior to the experiment. A preliminary report on this experiment without the key analyses was included in a talk and short paper for the 8th International Conference on Bioelectromagnetism (Horton et al. 2011).

Participants. Ten young adult volunteers (2 women, 8 men) between the ages of 21 and 29 yr participated in the study, but one was excluded from the analyses because of excessive artifact in the EEG. All participants were experienced EEG subjects with no reported hearing loss or history of neurological disorder. They completed the study in three to four sessions, typically spread over 2 wk.

Equipment. Participants were seated in a dim, sound-attenuated chamber facing a computer monitor and two loudspeakers (Fig. 1*A*). The loudspeakers flanked the monitor so that all three were 1.5 m from the subject and formed a 90° arc. FinalSound 400i electrostatic loudspeakers were used for stimulus presentation to ensure that the EEG was free of any speaker-generated electromagnetic interference, which can be produced by conventional magnet-driven loudspeakers. Participants held a keyboard in their lap, which they used to give behavioral responses. The experiment was controlled by a PC running MATLAB (The MathWorks, Natick, MA) and used functions from the Psychophysics Toolbox (Brainard 1997).

Stimuli. Each trial's stimulus consisted of independent left and right channels, each containing several sentence-length speech utterances. All sentences were taken from the TIMIT Acoustic-Phonetic Continuous Speech Corpus (Garofolo et al. 1993), which contains thousands of sentences spoken by both male and female speakers from a variety of dialect regions across the United States. Each channel was created by selecting random sentence waveforms from the corpus and concatenating them until the total length exceeded 22 s. Silent periods at the beginning or end of each sentence that were longer than 300 ms were reduced to 300 ms prior to concatenation, and sentences from the corpus were not reused within an experimental session. All sentences were normalized to have equal root mean square (RMS) power and then resampled from their original 16 kHz to the 48 kHz required to use the high-precision Audio Stream Input Output (ASIO) mode of the stimulus PC's sound card. Finally, the volumes of the loudspeakers were adjusted so that the mean intensity at the subject was 65 dB_{SPL}.

To induce steady-state responses in neural populations receiving auditory input, the left and right channels were amplitude-modulated via multiplication with sinusoids at 40 and 41 Hz, respectively (Fig. 1*B*). Those modulation frequencies elicit particularly robust ASSRs when used with tone or noise carriers (Picton et al. 2003), and similar modulations have been used to elicit ASSRs in normal, noisy, and reversed speech (Deng and Srinivasan 2010). Additionally, the modulation frequencies were sufficiently high to avoid interference with speech comprehension (Drullman et al. 1994a, 1994b; Miller and Licklider 1950). Subjects perceived the modulation as a rough quality in the speech, similar to the sound of someone talking through a fan. After hearing a few examples, they reported no difficulty in understanding the modulated speech.

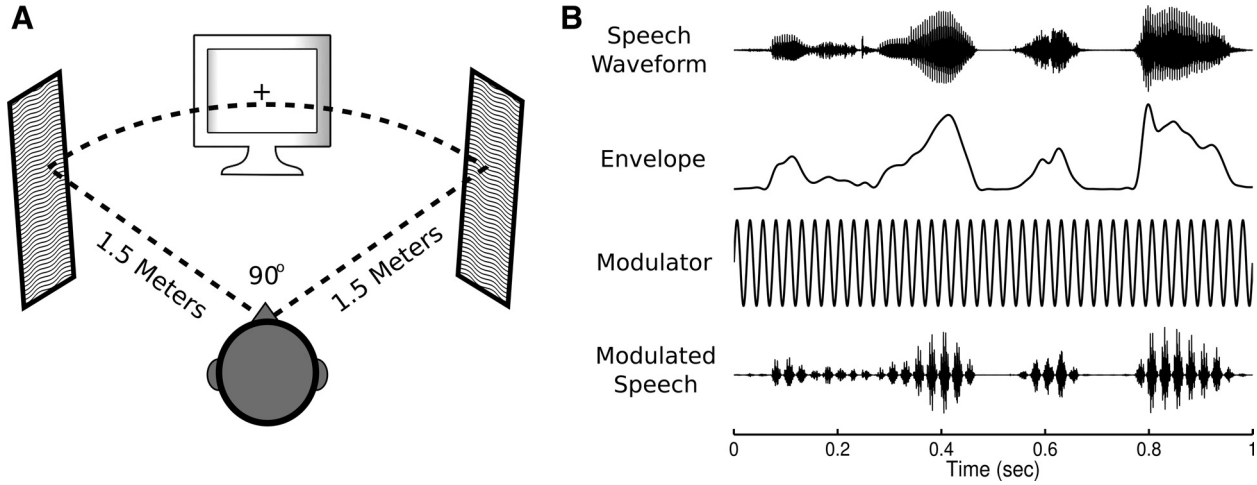

Fig. 1. *A*: the testing layout, showing the position of a computer monitor and 2 electrostatic loudspeakers relative to the subject. *B*, from *top* to *bottom*: the waveform for a partial sentence containing the words "the trips felt," its envelope filtered below 30 Hz, the 40-Hz modulator, and the stimulus waveform after the modulation has been applied.

Trial procedure. Each trial was randomly assigned to one of two conditions, "Attend Left" or "Attend Right." As depicted in Fig. 2, trials began with a written cue on the monitor, which displayed the condition for 2 s. The cue was then replaced with a small cross that subjects maintained fixation on for the duration of the trial. One second after the appearance of the fixation cross, the trial stimuli began playing through the loudspeakers. At their conclusion, the fixation cross disappeared and the transcript of one random sentence that had been used in the trial was displayed on the monitor. Participants were required to indicate with the "Y" and "N" keys on the keyboard whether the prompted sentence had originated from the cued speaker.

This particular behavioral task was chosen in order to encourage maximal suppression of the unattended speech. By deploying attention to the target speaker, and fully ignoring or "tuning out" the other, subjects could treat the behavioral task as a moderately challenging old/new judgment. The alternative strategy, monitoring both speakers during the trial and attempting to make a source judgment after seeing the prompt, proved to be extremely difficult because of the requisite memory load. In pilot testing, subjects performed at chance when asked to use this source monitoring strategy. Thus the former attend/ignore strategy was recommended to subjects during training. When debriefed, the subjects universally reported continuing to use that strategy throughout. In addition to requiring maximal suppression of the unattended speech, the behavioral task also encouraged participants to listen for meaning and commit the sentences to memory, as they did not know when the probe sentence would occur in the trial.

Participants were able to practice the behavioral task until they could maintain 80% accuracy over five trials. For all participants, this required $<$20 practice trials. They were encouraged to maintain this performance for the duration of the experiment, and they received feedback about their accuracy at the end of each experimental block. Participants completed 8 blocks, with 40 trials per block, for a grand total of 320 trials each, although one participant was only able to complete 6 blocks (240 trials) because of equipment failure.

Data collection and preprocessing. High-density EEG data were recorded during the experiment with a 128-channel electrode cap, with electrodes placed following the International 10/5 system (Oostenveld and Praamstra 2001) and all impedances kept below 10 kΩ. The cap, amplifier, and digitization software were all products of Advanced Neuro Technology. The EEG was sampled at 1,024 Hz with an online average reference and then subsequently imported into MATLAB for all further off-line analyses. Each channel was forward and backward filtered (to eliminate phase shift) with a Butterworth filter with a 1 to 50 Hz pass band and then downsampled to 256 Hz. On the occasions where a channel had been switched off during a recording session because of excessive noise, its data were replaced using a spline interpolation among neighboring channels. This interpolation was not needed for more than two channels on any subject (mean = 0.8 channels replaced per subject, SD = 0.78). For each trial, we extracted a 20-s epoch of EEG data beginning 1 s after the onset of the speech (Fig. 2). This slight delay ensured that analyses would not include the initial stimulus-onset responses, which are much larger than envelope-following responses and follow a stereotypical P1-N1-P2 pattern irrespective of the stimulus envelope (Abrams et al. 2008; Aiken and Picton 2008).

Trial data were screened by an experimenter, and those trials that contained large EEG artifacts that were not correctable, such as subject movements, were excluded from subsequent analyses. This screening resulted in an average of 16.6 trials excluded per subject

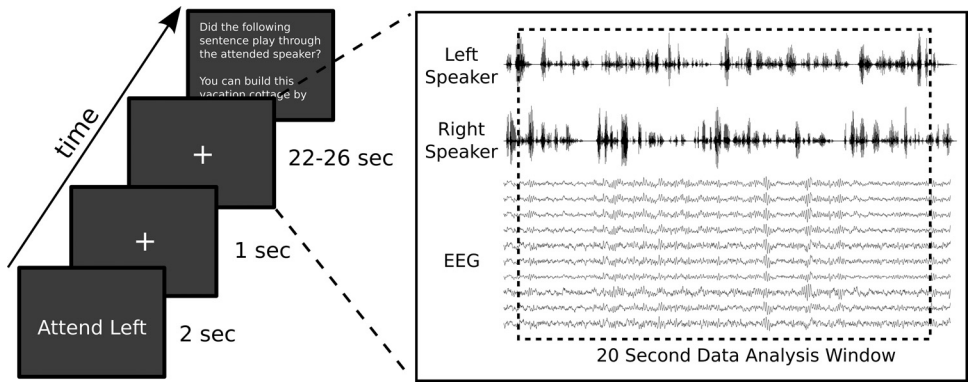

Fig. 2. *Left*: diagram showing the sequence of a trial. The attention cue appeared on screen for 2 s, followed by the appearance of a fixation cross, and then 1 s later the trial stimuli began. After the stimuli finished, subjects were asked to judge whether a prompted sentence was one of those played by the attended speaker. *Right*: detailed view of the listening period, illustrating that data analysis was restricted to a 20-s window that began 1 s after the onset of the stimuli.

(SD = 10.9). The remaining data for each participant were then submitted to the Infomax ICA algorithm, as implemented in standalone functions from the EEGLAB toolbox (Delorme and Makeig 2004). ICA attempts to separate spatially fixed and statistically independent sources that have been mixed together at the scalp (Lee et al. 1999). It has been shown to be effective for isolating and removing artifacts from EEG data (Jung et al. 2000) as well as identifying functionally independent brain sources or networks (Jung et al. 2001). To ensure that the ICA algorithm was presented a large amount of training data relative to the number of sources being estimated, we limited the input to the first 64 principal components of the data. Those 64 components accounted for an average of 98.4% (SD = 1.1%) of the variance in the EEG data for each subject, suggesting that the vast majority of the dynamics in the EEG were retained in this reduced form. The resulting ICA components were then reviewed to identify those that were attributable to blinks, muscle noise, electrical artifacts, or other sources of nonbrain activity. The components that clearly fit the topography, spectral content, and time series indicative of those artifacts were flagged as bad, and an artifact-corrected version of the data was created by projecting all nonartifact components back into channel space.

Speech envelopes. Speech envelopes were calculated by applying the Hilbert transform to the stimuli and then band-pass filtering from 2 to 30 Hz using forward and backward Butterworth filters. Finally, the envelopes were downsampled to 256 Hz to match the EEG. Some studies have elected to log-transform envelopes prior to filtering in an attempt to compensate for the logarithmic relationship between stimulus intensity and psychological percept intensity (Aiken and Picton 2008; Power et al. 2012). We performed all subsequent analyses using both standard and log-transformed envelopes and found no discernible differences between the results of each, which is not surprising given the high correlation between an envelope and its log transform after filters have been applied. For simplicity's sake, we only present results calculated from the standard envelopes.

Cross-correlation analysis. We measured phase-locked neural responses to the continuous speech stimuli by computing the cross-correlation function (Bendat and Piersol 1986) between each EEG channel and the acoustic envelopes of the attended and unattended speech. The cross-correlation function measures the similarity of two time series across a range of time lags. For discrete functions f and g, it is defined as

$$(f \star g)(n) = \sum_{m=-\infty}^{\infty} \frac{f[m]g[n+m]}{\sigma_f \sigma_g}$$

where σ_f and σ_g are the standard deviations of f and g. These functions appear similar to evoked potentials, with flat prestimulus baselines followed by a series of deflections away from zero. Peaks in the cross-correlation functions typically correspond to the latencies of well-known evoked potential components, since both measures rely on the underlying response characteristics of the auditory system. As in other methods that extract continuous responses to speech acoustics (Ding and Simon 2012a; Lalor et al. 2009), cross-correlation assumes a linear relationship between the stimulus envelope and the neural response. Intracranial recordings from nonprimary auditory cortex support that assumption for the low- and moderate-frequency modulations present in speech envelopes (Pasley et al. 2012).

For every trial, each channel of EEG was cross-correlated with the attended speech's envelope, the unattended speech's envelope, and a control envelope. As the attended and unattended envelopes were neither correlated with one another nor correlated with the control envelope, the responses to each could be estimated independently. The three cross-correlation functions for each channel were binned by condition, averaged over trials, and then averaged across subjects. The control envelope for each trial was the envelope of a different trial that was presented to that subject, selected at random. This control was useful because it shared all of the spectral and temporal characteristics of the attended and unattended envelopes but was unrelated to that particular trial's stimuli. Therefore, any nonzero values in the control cross-correlation functions were due purely to chance. We collapsed the measured values in the control cross-correlation function across time and channel, forming an approximately Gaussian distribution (Fig. 3*B*). This distribution was then used to compute a 99% confidence interval for the null hypothesis that (at a given latency) no correlation existed between the control envelope and the EEG. Cross-correlation values between the attended or unattended envelopes and EEG channels greater than the 99.5th percentile of this control distribution, or less than the 0.5th percentile, were deemed to be significantly nonzero (2-tailed $P < 0.01$).

Even with a fairly conservative threshold for significance, the 128 (channels) × 205 (delays) comparisons involved in testing each function necessitated a procedure to address the multiple comparisons problem. If the null hypothesis were true, and no phase-locked relationship existed between an EEG channel's activity and one of the stimuli's envelopes, we would still record a significant correlation value 1% of the time because of chance variation. We chose to control the false discovery rate (FDR), the proportion of significant results that are expected to be incorrect rejections of the null (Benjamini and Yekutieli 2001). With a large amount of comparisons, the FDR seeks to strike a reasonable balance between statistical power and type I error. To keep the ratio of incorrect rejections to all rejections less than α for m independent (or positively dependent) tests, the procedure first ranks the observed P values in ascending order:

$$P_{(1)} \leq P_{(2)} \leq \ldots \leq P_{(m)}$$

Then one finds the largest index k so that

$$P_{(k)} \leq \frac{k}{m}\alpha$$

The null is then rejected for the tests associated with the P values with indices from 1 to k. We used this procedure to adjust our significance threshold so that the FDR remained at 1% in the attended and unattended cross-correlation functions.

Steady-state analysis. Trial data were transformed into the frequency domain with MATLAB's fast Fourier transform. Since the individual trials were 20 s long and contained an integer number of cycles of each modulator, the transformed data contained frequency bins that were 0.05 Hz wide and centered on the modulation frequencies. The Fourier coefficients that corresponded to the amplitude modulations at 40 and 41 Hz were binned by condition, averaged over trials, and then averaged across subjects. Statistical comparison of the amplitudes of the steady-state responses were conducted with a repeated-measures analysis of variance (RMANOVA) with condition (Attend Left, Attend Right), stimulus location (left, right), and electrode site (Fz, Cz, O1, etc.) as factors.

ICA decomposition and clustering. Speech processing involves multiple structures within the cortex (Hickok and Poeppel 2007). The cross-correlation and steady-state responses reflect the summed activity of all of those structures that responded to the speech stimuli. Determining the location and relative contribution of individual brain sources from this summed response can be extremely difficult, as their responses overlap in time and have nonorthogonal scalp distributions. To better describe the responses of individual areas involved in speech processing, we decomposed each subject's data using ICA. The ICA process can leverage the statistical properties of the entire data set to separate the activity of individual brain areas or networks, provided that those areas/networks are not always jointly active (Jung et al. 2001). Each subject's artifact-corrected data were submitted to the Infomax ICA algorithm, with the condition that enough components were kept to account for 95% of the variance in the EEG. This resulted in 12–18 components per subject with scalp distributions typical of the large synchronous populations that generate the bulk of EEG data (Nunez and Srinivasan 2006).

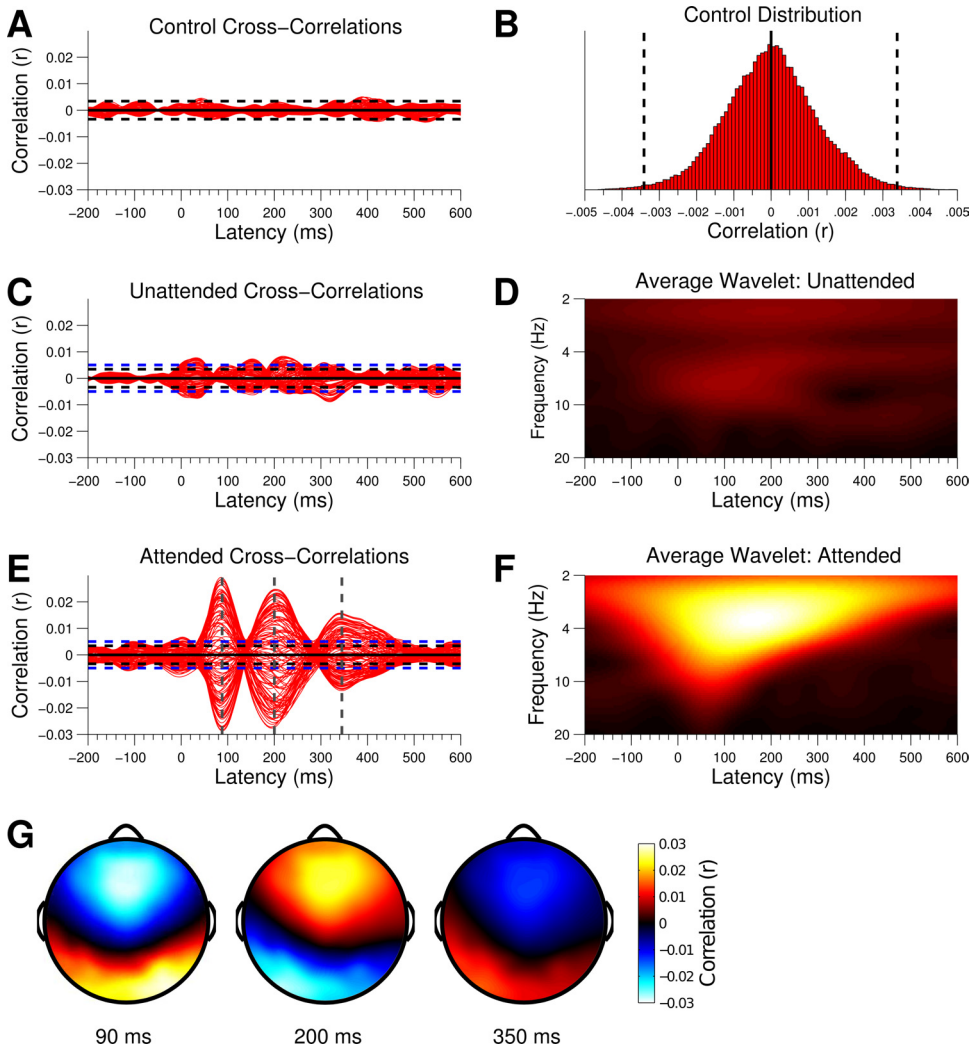

Fig. 3. *A*: cross-correlation functions between the control envelopes and EEG channels. Each trace represents an individual channel. *B*: control cross-correlations were collapsed across channel and time to form a null distribution. The 0.5th and 99.5th percentiles were set as the threshold of significance for a single test, indicated by black dashed lines. *C* and *E*: unattended and attended cross-correlation functions. Blue dashed lines indicate the threshold for significance after adjustment for multiple comparisons. *D* and *F*: wavelet spectrograms of the cross-correlation functions, averaged over channels, indicating the frequencies at which phase-locking occurred. *G*: scalp topographies for the peaks in the attended cross-correlation function. Warm colors denote correlations with positive potentials, while cool colors denote correlations with negative potentials.

With each subject's data now divided into a set of ICA components, we then needed a way to cluster (match) components across subjects. This process can be difficult for a variety of reasons (Onton and Makeig 2006). First, the scalp projection of any given brain source can vary between subjects because of differences in brain shapes and volume conduction through the head. Second, it is not guaranteed that each subject's data will decompose into an equivalent set of components. Furthermore, multiple components within a single subject can share similar topographies, provided their activity remains sufficiently independent to be identified as distinct functional networks. With these complications, the most appropriate clustering scheme depends on the specific goals of the clustering (Onton and Makeig 2006). We sought to maximize the performance of the clustering for those components that phase-locked to speech envelopes and contributed to the channel-space cross-correlation functions. To that end, we sought to eliminate ICA components that did not respond to the experimental stimuli. Thus we set the criterion that to be included in the clustering process an ICA component had to show significant ASSRs, defined as having greater power at the stimulus modulation frequencies than the 99th percentile of the surrounding 100 frequency bins. As steady-state responses typically propagate into areas downstream from where they are generated, this criterion retained temporal, parietal, and frontal sources that were likely to be involved in audition and attention, while excluding components with scalp topographies and spectra indicative of motor, somatosensory, and visual cortex sources. The excluded components were subsequently examined, and we confirmed that none displayed evidence of cross-correlation with the speech envelopes. For those components that fit the criterion, we then used the normalized electrode weights (scalp topography) and the normalized power spectra from 0 to 50 Hz to cluster the data.

We used a standard *k*-means clustering algorithm that minimized the sum of squared Euclidean distances of the components from their cluster centroids. For *k*-means clustering, one must choose the number of clusters to fit, as opposed to it being determined from the data. We clustered the data with multiple values of *k* ranging from 4 to 9. For each *k*, we repeated the clustering 1,000 times and kept the best fit for further evaluation. Since the total distance measure will always shrink with each additional cluster, a balance must be struck between the number of clusters and goodness of fit. We used a common metric for gauging the optimal number of clusters that seeks to maximize the ratio of between-cluster sum of squared distances to within-cluster sums of squared distances (Caliński and Harabasz 1974).

Analyses of clustered data. As in the channel-space data, we used cross-correlation to extract neural responses to the speech stimuli that were phase-locked to their envelopes. Additionally, we computed wavelet spectrograms of each cross-correlation function to describe their time-frequency content. For this, we used an array of complex-valued continuous Morlet wavelet transforms with parameters that stressed time resolution over frequency resolution.

With far fewer clusters to examine than individual EEG channels, we were able to use bootstraps (Efron and Tibshirani 1993) to build 99% confidence intervals for the cross-correlation functions rather than simply setting thresholds for significantly nonzero responses.

Bootstraps are computationally prohibitive but are advantageous in that they do not require assumptions about the shape of the underlying distribution. Using MATLAB's bootstrapping functions, we performed 5,000 different resamplings of the data. These consisted of n trials randomly selected with replacement, where n was equal to the total number of trials in the data. Each resampling of the data produced an estimate of the true cross-correlation function, and the distribution of those estimates was used to determine the bias in the estimator and subsequently a bias-corrected confidence interval for the true cross-correlations. As in the channel-space cross-correlation analyses, the confidence intervals were then adjusted for multiple comparisons to maintain a familywise confidence of 99%. Similar bootstrap procedures were also used to calculate 99% confidence intervals (after correcting for multiple comparisons) for the wavelet spectrograms of the cross-correlation functions and the ASSRs.

RESULTS

Behavior. The participants performed well on the sentence recognition task, given its challenge, with a mean accuracy of 82.45% (SD = 4.85%). Participants reported in the debriefing that most of their errors were due to memory constraints rather than lapses in attention. While that may have been true, the behavioral task was not capable of discriminating between those two sources of error.

Envelope-EEG channel cross-correlations. We used cross-correlation to extract phase-locked neural responses to the speech envelopes. Grand-averaged cross-correlation functions for control, attended, and unattended envelopes are plotted in Fig. 3, with each trace representing an individual EEG channel. As expected, the control envelopes (Fig. 3*A*) showed no systematic relationship to the EEG, with correlations fluctuating around zero. The distribution of correlations, collapsed across channels and latencies, was approximately Gaussian (Fig. 3*B*). We used the 0.5th and 99.5th percentiles of that distribution (indicated by black dashed lines in Fig. 3*B*) to determine the maximum amount of correlation that we expected to occur by chance (before multiple comparisons procedures) in the attended and unattended cross-correlation functions.

The attended cross correlation functions appear in Fig. 3*E*; blue dashed lines indicate the level of the smallest significant correlation value after adjusting for multiple comparisons. We found robust phase-locked responses to the attended speech's envelope, with highly significant peaks in correlation values at 90-, 200-, and 350-ms latencies—corresponding to the N1, P2, and N2 components from the auditory evoked potential literature (Picton et al. 1974). In looking at the scalp distributions of the peaks (Fig. 3*G*), the response at 90 ms was somewhat right-lateralized, while those at 200 and 350 ms were conversely left-lateralized. The time-frequency representation of the cross-correlation functions (Fig. 3*F*) indicated that entrainment of endogenous oscillations had occurred primarily in the low theta to delta band, with a gradual decrease in frequency over time.

We could see immediately from the unattended cross-correlation functions (Fig. 3*C*) that attention had a major impact on phase-locking to speech envelopes. The large peaks in correlation that we saw in the attended cross-correlations were all but absent in the unattended case, with fewer channels crossing the threshold for significance. Additionally, the latencies at which the strongest correlations were recorded for the unattended functions did not match the corresponding latencies in the attended functions. We would expect the peaks to occur at the same latencies if the attended and unattended responses were simply scaled versions of one another. Overall, while the unattended cross-correlation functions clearly contained structure that was not present in the control functions, the small amount of signal relative to noise made it difficult to form a detailed profile of the unattended response. The time-frequency representation of the unattended cross-correlations (Fig. 3*D*) suggested that what phase-locking occurred resided mostly in the theta band.

Auditory steady-state responses. The 40- and 41-Hz-amplitude modulations elicited robust ASSRs in all participants at the exact frequencies of modulation. The grand-averaged amplitude spectrum of midfrontal channel Fz (Fig. 4*A*) illustrates the size of the ASSRs relative to the full amplitude spectrum of the EEG. The ASSRs were small but still clearly visible. As the steady-state responses were phase-locked across all trials, we could also average the complex Fourier coefficients instead of just the amplitudes. This operation is equivalent to averaging the trials in the time domain and then Fourier transforming that average. By doing this, activity that was phase-locked across trials was preserved, while all other activity averaged toward zero. Figure 4*B* illustrates the boost in signal-to-noise ratio that resulted from this form of averaging. All further plots and statistical analyses of the ASSRs used those complex-averaged data.

Figure 4*C* depicts topographic polar plots of the steady-state responses. For each channel, the length of the line represents the amplitude of the response and the angle represents the phase. The scalp distributions were typical for ASSRs, with both frontal and occipital maxima that were ~180° out of phase. ASSRs are commonly thought to result from a combination of brain stem and auditory cortical sources, whose responses overlap to form this frontally peaked distribution (Herdman et al. 2002). When comparing the ASSRs between the attended and unattended stimuli on each side, we observed no differences in phase or latency. In all respects, ASSRs elicited by attended stimuli appeared identical to ASSRs elicited by unattended stimuli. Although we observed no effects of attention, we did see some differences between responses to stimuli on the left and right. The ASSR amplitudes were slightly larger for stimuli on the right, and their distribution across the scalp was subtly different. In addition, the absolute phases of the left and right ASSRs were slightly different because of the 1-Hz difference in their modulation frequencies.

The results of the ANOVA reinforced these observations. Steady-state amplitudes were not modulated by attention ($F_{1,8} < 0.01$, not significant). ASSRs to stimuli on the right had larger amplitudes than ASSRs to stimuli on the left ($F_{1,8} = 313.1$, $P < 0.001$). The differences between the topographies of ASSRs to the left and right stimuli resulted in a significant interaction between stimulus side and electrode location ($F_{1,127} = 3.77$, $P < 0.001$). Finally, the ANOVA also revealed that the side of attention did not alter overall ASSR amplitudes ($F_{1,8} = 6.03$, $P = 0.14$).

ICA component clustering. For each value of k from 4 to 9, we used the best fit from 1,000 iterations of the k-means clustering algorithm to evaluate the goodness of fit. We found that six clusters maximized Caliński and Harabasz's (1974) criterion for the optimal number of clusters. Solutions for more than six produced clusters with only one ICA component assigned to them, indicating that only six major trends could be

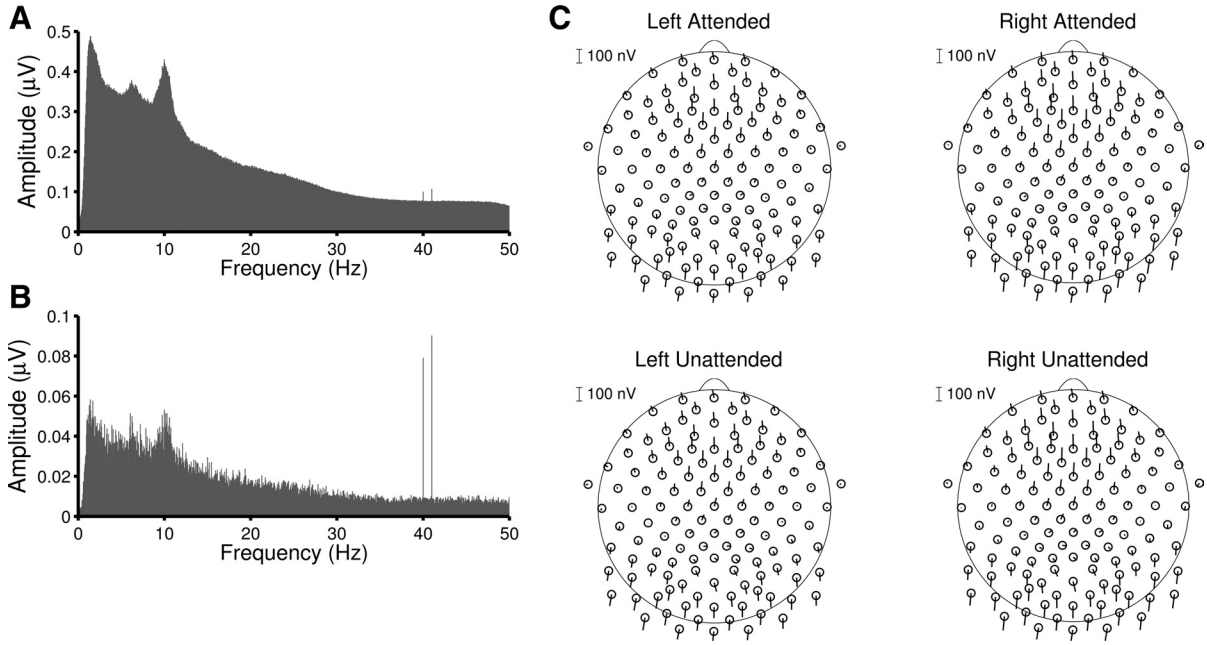

Fig. 4. *A*: grand-averaged amplitude spectrum of frontal channel Fz, showing that the steady-state responses at 40 and 41 Hz are small but visible in individual trials. *B*: the complex Fourier coefficients for each trial can be averaged to reduce the contribution of all activity that is not phase-locked across trials. The steady-state responses are preserved, while all other activity falls off. *C*: polar plots showing the amplitude (length) and phase (angle) of the steady-state responses. Channels located outside the ring are located below the plane formed between the nasion and occipital channels O1 and O2.

identified across the set of ICA components and the algorithm was being forced to pick out the least representative components to populate the additional clusters. The scalp topographies and power spectra of the six-cluster solution are shown in Fig. 5, *A* and *B*, respectively. For each cluster, we used EEGLAB's dipole fitting functions to estimate the location and orientation of the equivalent dipole source that best accounted for its scalp topography (Delorme and Makeig 2004). Although equivalent dipoles are admittedly oversimplified models for brain sources, they produce very good fits for the scalp maps produced by ICA and other blind source separation techniques (Delorme et al. 2012). We viewed the fits simply as a rough estimate of the centroid of the brain sources, with full knowledge that other source configurations could produce identical scalp topographies. Clusters with lateralized topographies (*1*, *2*, *4*, and *5*) were fit with a single dipole. Those with topographies that were focal over the midline (*3* and *6*) were fit with both single and symmetric dipole sources.

The dipole fits for *clusters 1* and *2* were located in the left and right posterior temporal lobes, near the superior temporal sulci (STS). Dipole fits for *clusters 4* and *5* were located more anteriorly in the temporal lobe, closer to primary auditory areas. These dipole fits have good face validity, in that these are cortical areas known to be involved in speech processing (Hickok and Poeppel 2007). *Cluster 6* was well-fit with either one or two dipoles located near the midline in the parietal lobe, which agreed with the strong alpha rhythm in its power spectrum. Finally, the underlying sources for *cluster 3* were far less certain. The best single dipole fit was located under the frontal lobe, right above the thalamus, while the fit for symmetric dipoles was located shallower in the frontal lobes. Although auditory responses are often maximal in frontal electrodes, they have been shown to originate primarily from a combination of sources in the brain stem and the primary auditory regions on the superior surface of the temporal lobe (Herdman et al. 2002; Vaughan and Ritter 1970). Accordingly, we found that placing dipoles in either of those locations could also account for *cluster 3*'s scalp topography with low residual variance. Therefore, we felt that it was most appropriate to assume that *cluster 3* could be sensitive to any and all of those brain structures.

Envelope-cluster cross-correlations. We cross-correlated the clustered data with the attended, unattended, and control speech envelopes. Each cluster's attended (red) and unattended (gray) cross-correlation functions appear in Fig. 5*C*. The shaded regions indicate 99% confidence intervals for the means after correction for bias and multiple comparisons. The control cross-correlation functions never differed significantly from zero and are not plotted. The attended and unattended cross-correlation functions in *clusters 4*, *5*, and *6* were also nearly flat, indicating that activity in those brain areas did not significantly phase-lock to the envelopes of attended or unattended speech.

In contrast, *clusters 1*, *2*, and *3* showed large responses with strong differences between attended and unattended speech. We projected just those three clusters back into channel space and recovered attended and unattended cross-correlation functions that were indistinguishable at a glance from those in Fig. 3, confirming our suspicions that the channel-space cross-correlations were actually the sum of temporally overlapping activity stemming from these three brain areas. The time-frequency representations of the cross-correlation functions appear in Fig. 5, *D* and *E*. Dashed lines in the attended wavelet spectrograms indicate time-frequency areas where the bootstrap analyses indicated that the attended responses were significantly greater than the unattended responses.

The attended cross-correlation functions in *clusters 1* through *3* peaked at the same latencies (90, 200, and 350 ms) that were observed in the channel-space cross correlations. The overall shape of each of those cluster's attended response was

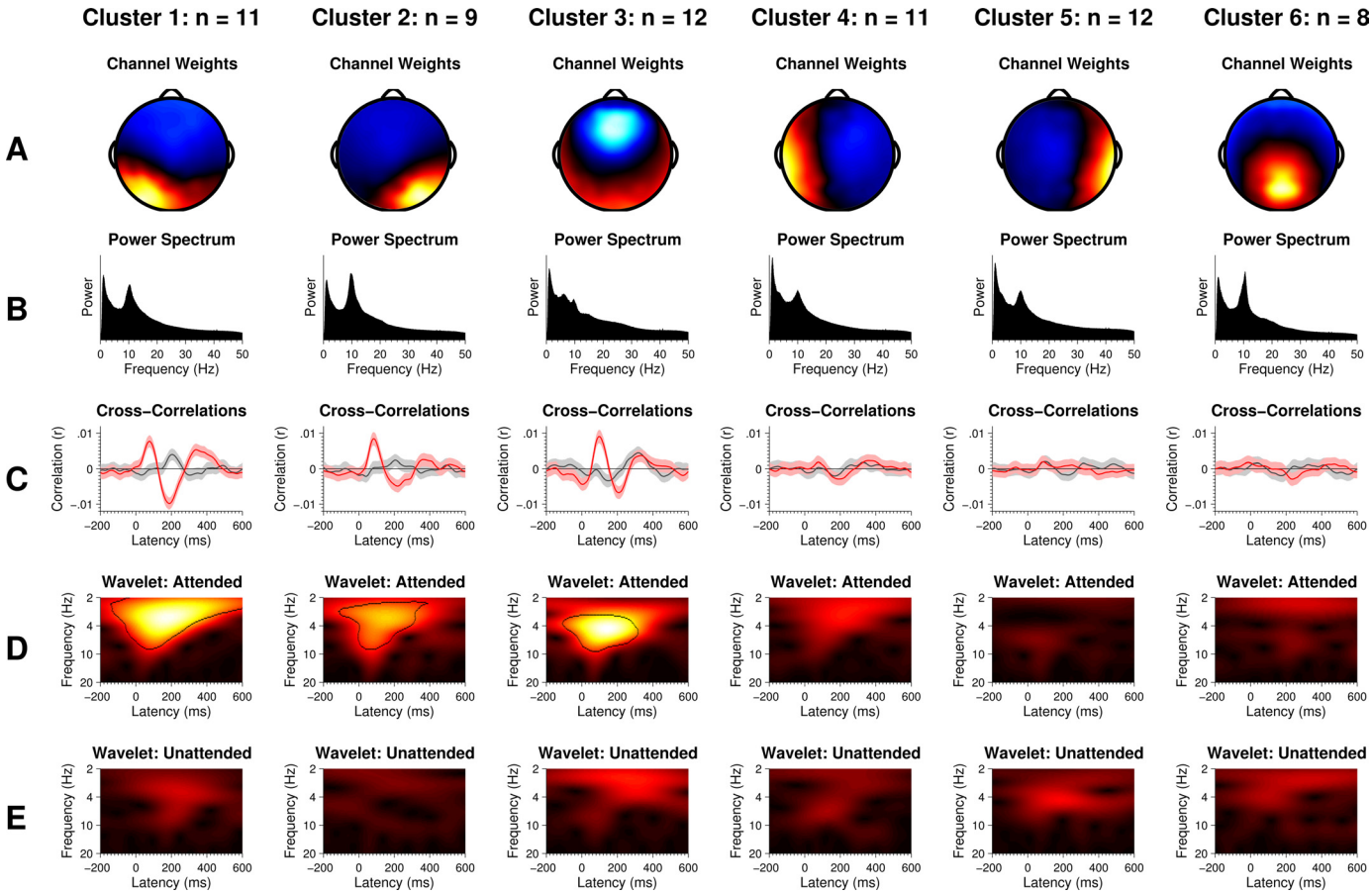

Fig. 5. *A*: average channel weights for each cluster. Warm colors denote positive weights, cool colors denote negative weights, and black denotes zero weights. *B*: average power spectrum for each cluster. *C*: average cross-correlation functions for each cluster with shaded areas denoting 99% confidence intervals. Red and gray lines correspond to the attended and unattended cross-correlations, respectively. *D* and *E*: wavelet-spectrogram representations of the cross-correlation functions. Regions enclosed by dashed lines indicate time-frequency regions where the attended cross-correlation function has significantly more power than the unattended function.

similar (i.e., a positive, a negative, then another positive deflection from zero), but the strength of each peak varied. *Clusters 1* and *3* had robust responses at all three latencies. *Cluster 2* had a strong response at 90 ms but greatly reduced responses at 200 and 350 ms. *Cluster 3* also contained a significant response at a 15-ms latency, which is too fast for a cortical response and more likely reflected the brain stem response to speech (Chandrasekaran and Kraus 2010). In looking at the time-frequency representations of the attended responses (Fig. 5*D*), we observed that the peak frequency content of *clusters 1* and *2* fell at the edge of the delta and theta bands, while that of *cluster 3* fell slightly higher in the low theta band. That could indicate a difference in which frequencies in the envelope were preferably tracked by each cluster's source. Alternatively, the frequency content of *clusters 1* and *2* could appear lower if their responses were less consistent in latency, since temporal jitter tends to smooth out estimations of evoked responses.

Turning to the unattended cross-correlations, we observed across the board that responses were smaller than those to attended speech. However, the unattended and attended cross-correlation functions were sufficiently different to make it obvious that they were not all just scaled versions of one another as reported in previous studies. The unattended cross-correlation functions for *clusters 1* and *2* were similar in shape, but the magnitude of the unattended response in *cluster 2* was somewhat weaker. In these clusters, we observed two primary differences from their attended cross-correlations. First, the initial peaks that occurred at 90 ms in the attended functions were completely absent in the unattended functions. Second, from 150- to 450-ms latencies, the signs of the correlations were opposite to those in the attended functions. Thus, across those latencies, the brain areas corresponding to *clusters 1* and *2* directly encoded the envelope of speech if it was attended or encoded the inverse of the envelope if it was unattended, consistent with the hypothesized mechanism of suppression through entrainment. *Cluster 3* did not show the sign inversion seen in the previous two clusters. The shape resembled a scaled-down version of the attended response, in that they both trended up or down at similar times. The unattended response begins smaller, but by 300-ms latency was indistinguishable from the attended response. The time-frequency representations of the unattended responses (Fig. 5*E*) were mostly uninformative because of the small sizes of their responses, but what power they contained fell into the delta and theta bands.

Auditory steady-state responses in clusters. Using the confidence intervals produced by the bootstrap analyses, we found that the ASSRs were significantly larger in *cluster 3* than in any other cluster, consistent with the close match between its topography and that of the ASSR we obtained from the

pre-ICA EEG. However, the ASSRs were not modulated by attention in any of the six clusters.

DISCUSSION

Summary. In the present study, we created a "cocktail party" scenario wherein subjects needed to attend to one speaker while suppressing a competing speaker. We found three distinct brain areas/networks that phase-locked to the acoustic envelopes of attended and unattended speech. In all three, the responses to unattended speech were weaker overall than those to attended speech. This was especially pronounced for the earliest cortical responses peaking around 90 ms. We also found a previously unreported effect, where sources in the posterior temporal cortices encoded the envelope of attended speech at the same latency at which they encoded the inverse of the unattended speech's envelope. Additionally, we evoked ASSRs by modulating the amplitudes of the speech stimuli at 40 and 41 Hz and found that responses did not differ with attention.

Phase-locking to attended speech. We observed robust phase-locked responses to the envelopes of attended speech streams. Those responses fit the timing and scalp distribution of the classic N1-P2-N2 auditory evoked potential components (Picton et al. 1974) and largely reproduced envelope responses that have been reported in comparable studies (Abrams et al. 2008; Aiken and Picton 2008; Ding and Simon 2012a; Lalor and Foxe 2010). The late peak (N2) in our attended cross-correlation function has not always been present in other studies. This component is sensitive to several cognitive functions including attention, novelty detection, and cognitive control (Folstein and Van Petten 2008) and thus may be more visible in the present study because of the particular set of cognitive demands in our behavioral task.

In addition to replicating previous work, our ICA decomposition revealed that the phase-locked responses that we (and previous studies) recorded at the scalp electrodes were not generated by a single source within the brain. Rather, we found evidence of three independent brain sources/networks that phase-locked to speech envelopes. These responses shared similar time courses, yet their relative strengths varied across latency. These differences, previously unobserved, have important implications regarding the hemispheric lateralization of function. Several models of auditory processing have proposed that the left and right auditory cortices have different responsibilities (Friederici and Alter 2004; Poeppel 2003; Zatorre et al. 2002). According to the "asymmetric sampling in time" (AST) hypothesis, for example, speech processing is a bilateral effort, but each hemisphere preferentially processes certain timescales (Poeppel 2003). The left hemisphere is thought to be focused on very short timescales, needed to discriminate place of articulation and other fast temporal features, while the right hemisphere preferentially follows the slower temporal features such as the envelope. Most studies report better representations of the acoustic envelopes in the right hemisphere (Abrams et al. 2008; Ding and Simon 2012a; Kerlin et al. 2010; Luo and Poeppel 2007), but some have found no difference or the opposite trend (Aiken and Picton 2008; Millman et al. 2011; Peelle et al. 2012).

Our results show evidence of both right and left lateralization for envelope tracking, depending on the time window of interest. Thus it would seem inappropriate to declare one hemisphere as dominant for envelope processing in general. Rather, lateralization of function is better described separately for early and late cortical responses. The earliest cortical response to speech is thought to relate to its spectrotemporal analysis (Hickok and Poeppel 2007). In our data, this response (at 90 ms) was slightly stronger in the right hemisphere source, which lends some support for the AST hypothesis. In contrast, the later cortical responses were very strongly left-lateralized. Left hemisphere cortical structures are known to dominate lexical/semantic stages of speech processing (Hickok and Poeppel 2007) and are preferentially activated by connected (as compared with single syllables or words) speech (Peelle 2012). Thus prolonged maintenance of the acoustic envelope in the left hemisphere may be involved in the mapping from spectrotemporal features to lexical/semantic meaning. This view is further reinforced by the observation that envelope tracking in the left hemisphere is reduced for unintelligible speech (Peelle et al. 2012).

Phase-locking to competing speech. We found significant phase-locked responses to the competing speech within the same three clusters that responded to the attended speech. The use of ICA was particularly helpful in describing the responses to the competing speech, as they destructively interfered at the scalp electrodes and therefore were masked in the channel-space cross-correlation functions. It is likely that the effects reported in previous studies also reflect the combined activity of these three sources and thus may have obscured the full range of attentional effects in the present study. In all three of the clusters that recorded phase-locked responses, the responses to unattended speech were weaker overall than those to attended speech. This was particularly true for the initial cortical responses, which were almost entirely absent in the unattended cross-correlation functions. This mostly agrees with previous studies that have reported suppression of the unattended responses (Ding and Simon 2012a; Power et al. 2011, 2012) or preferential encoding of the attended speech (Kerlin et al. 2010; Mesgarani and Chang 2012). However, the degree to which the unattended responses were diminished relative to the attended responses was greater in the present study than in any previous work. This does not directly contradict previous studies, as the behavioral task in the present study was very demanding and was specifically designed to require more thorough suppression of the competing speech. Previous studies typically required subjects to finish attended sentences or answer questions regarding an attended passage, which would not necessarily be disrupted by additional information from the unattended side. However, in our task the failure to suppress or "tune out" the unattended speech was disastrous to the subjects' performance. Thus our findings could be viewed as a more powerful example of the same fundamental effect.

Suppression of competing speech through entrainment of cortical oscillations. We hypothesized that attention networks could suppress a competing speech stream by encouraging auditory populations to phase-lock to the inverse of that stream's envelope. We found sources in the left and right posterior temporal cortices that strongly encoded the envelope of attended speech at 200-ms latency yet encoded the inverse of the unattended speech's envelope at that same latency. As changes in gain alone should not be able to invert the sign of

the correlation, we believe the hypothesized mechanism of entrainment is the most parsimonious explanation for the observed inverse envelope encoding. Thus these data constitute the first evidence that we are aware of in which attention networks have manipulated the phase of slow neuronal oscillations to suppress a competing rhythmic stimulus.

While this helps to understand the neural processing of speech during "cocktail party" scenarios, many questions remain about how this process takes place and what purpose it serves. First, it remains uncertain whether this entrainment mechanism is capable of enhancing an attended speaker and suppressing a competing speaker at the same time. We saw evidence of both in the averaged cross-correlations, but that does not mean that they occurred concurrently. Attention networks may have dynamically switched from enhancing the target stream to suppressing the competing stream, depending on which was the most effective strategy at that moment. We are limited to observations from the averaged data, as we cannot reconstruct speech envelopes from the raw EEG of single trials with sufficient accuracy. However, recent studies using implanted electrodes have demonstrated excellent reconstruction of stimulus envelopes (Mesgarani et al. 2009; Mesgarani and Chang 2012; Pasley et al. 2012) and may clarify this point in the future.

Second, it remains uncertain whether the primary function of phase-locking is, in fact, to align syllable arrivals with periods of maximum (or minimum for competing) neural excitability. The beginnings of syllables do not necessarily contain the most critical information for comprehension. The ends of (and transitions between) syllables can also supply important information, yet this mechanism would effectively suppress that information. Further research is needed to clarify why phase-locking to the beginnings of syllables is optimal. Perhaps changes in neuronal excitability are not actually the primary goal of entrainment. Instead, entrainment may primarily underscore the segmentation of speech. It has been suggested that the parallelized processing of speech at multiple timescales (i.e., phoneme, syllable, word, sentence, narrative) requires coordination across several frequency bands of cortical oscillations (Ghitza 2011; Ghitza and Greenberg 2009; Giraud and Poeppel 2012). The envelope can provide the boundaries for segmentation at each of these timescales, and entrainment may be required to maintain synchronization across frequency bands. If true, entrainment to the inverse of the competing speech's envelope may act to undermine the proper segmentation of the competing speech stream.

Invariance of ASSRs to attention. We attempted to measure changes in gain resulting from attention by comparing ASSRs elicited by the attended and unattended speech. As in a recent similar study (Ding and Simon 2012a), we found that attended and unattended speech elicited identical ASSRs. Interpretation of this result is difficult, as there are many possible reasons why the ASSRs were not significantly modulated.

First, the ASSRs might have been identical when attended or unattended because they are primarily generated in the midbrain and auditory core (Herdman et al. 2002). As the differences in entrainment were observed in later stages of the auditory pathway, the changes in gain may also have been limited to those later areas and thus unobservable with an ASSR. Even if we assume that the differences in entrainment had altered stimulus gain at the level of the auditory core through top-down influence, the changes in excitability due to the phase of delta and theta oscillations are most noticeable for near-threshold stimuli (Lakatos et al. 2005). Since the modulations that drive ASSRs are quite salient and by definition highly entraining, they may be insensitive to subtle changes in excitability.

Additionally, we may not have observed effects of attention in the ASSRs because the 40- and 41-Hz-amplitude modulations were both being suppressed (or ignored) as distracting features in the speech stimuli. The modulations carried no linguistic meaning, yet they overlapped in timescale with modulations that determine phonemic-scale features such as place of articulation (Poeppel 2003). Thus it may have been beneficial to suppress the constant modulations in order to better judge transient modulations of the same timescale. This account has some empirical support; modulated speech and speech-in-noise produce smaller ASSRs than reversed speech and nonspeech carriers (Deng and Srinivasan 2010), indicating that people only suppress these uninformative modulations when processing stimuli for linguistic meaning.

Although steady-state responses to visual stimuli are strongly modulated by attention to location and feature (Andersen et al. 2011), the literature is heavily divided regarding the effects of attention on ASSRs. The ability to observe attentional effects in ASSRs seems to greatly depend upon stimulus characteristics, task demands, and level of mental arousal (Bidet-Caulet et al. 2007; Lazzouni et al. 2010; Linden et al. 1987; Müller et al. 2009; Ross et al. 2004; Skosnik et al. 2007). In the future, a different type of auditory probe may be more appropriate in measuring changes in gain when using speech stimuli, such as chirps or oddball sounds inserted into the speech streams. Comparison of the subsequent evoked responses may be able to better quantify gain changes in later cortical stages.

Conclusions. This study provides evidence that phase-entrainment mechanisms are used by attention networks for both enhancement and suppression of speech streams. These mechanisms are effective for speech largely because the frequencies of nested endogenous neural activity (delta, theta, and gamma) represent the dominant timescales of speech information. However, it remains unclear if these preexisting neural constraints guided the development of speech structure, or if the proliferation of complex speech selected for this neural organization. If it is the latter, then each sensory system may contain oscillatory hierarchies that reflect their typical rates of stimulation.

GRANTS

This work was supported by National Institute of Mental Health Grant 2R01-MH-68004 and Army Research Office Grant ARO 54228-LS-MUR.

DISCLOSURES

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

AUTHOR CONTRIBUTIONS

Author contributions: C.H., M.D., and R.S. conception and design of research; C.H. performed experiments; C.H. analyzed data; C.H., M.D., and R.S. interpreted results of experiments; C.H. prepared figures; C.H. drafted manuscript; C.H., M.D., and R.S. edited and revised manuscript; C.H., M.D., and R.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 SS, 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.

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

Alcaini M, Giard M, Echallier J, Pernier J. Selective auditory attention effects in tonotopically organized cortical areas: a topographic ERP study. *Hum Brain Mapp* 2: 159–169, 1995.

Andersen SK, Fuchs S, Müller MM. Effects of feature-selective and spatial attention at different stages of visual processing. *J Cogn Neurosci* 23: 238–246, 2011.

Bendat JS, Piersol AG. *Random Data: Analysis and Measurement Procedures*. New York: Wiley, 1986.

Benjamini Y, Yekutieli D. The control of the false discovery rate in multiple testing under dependency. *Ann Stat* 29: 1165–1188, 2001.

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.

Brainard DH. The Psychophysics Toolbox. *Spat Vis* 10: 433–436, 1997.

Bridwell DA, Srinivasan R. Distinct attention networks for feature enhancement and suppression in vision. *Psychol Sci* 23: 1151–1158, 2012.

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

Caliński T, Harabasz J. A dendrite method for cluster analysis. *Commun Stat* 3: 1–27, 1974.

Canolty RT, Edwards E, Dalal SS, Soltani M, Nagarajan SS, Berger MS, Barbaro NM, Knight RT. High gamma power is phase-locked to theta oscillations in human neocortex. *Science* 313: 1626–1628, 2006.

Chandrasekaran B, Kraus N. The scalp-recorded brainstem response to speech: neural origins and plasticity. *Psychophysiology* 47: 236–246, 2010.

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

Da Costa S, Van der Zwaag W, Miller LM, Clarke S, Saenz M. Tuning in to sound: frequency-selective attentional filter in human primary auditory cortex. *J Neurosci* 33: 1858–1863, 2013.

Darwin CJ, Hukin RW. Auditory objects of attention: the role of interaural time differences. *J Exp Psychol Hum Percept Perform* 25: 617–629, 1999.

Delorme A, Makeig S. EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *J Neurosci Methods* 134: 9–21, 2004.

Delorme A, Palmer J, Onton J, Oostenveld R, Makeig S. Independent EEG sources are dipolar. *PloS One* 7: e30135, 2012.

Deng S, Srinivasan R. Semantic and acoustic analysis of speech by functional networks with distinct time scales. *Brain Res* 1346: 132–144, 2010.

Desimone R, Duncan J. Neural mechanisms of selective visual attention. *Annu Rev Neurosci* 18: 193–222, 1995.

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

Ding N, Simon JZ. Emergence of neural encoding of auditory objects while listening to competing speakers. *Proc Natl Acad Sci USA* 109: 11854–11859, 2012b.

Drullman R, Festen JM, Plomp R. Effect of reducing slow temporal modulations on speech reception. *J Acoust Soc Am* 95: 2670–2680, 1994a.

Drullman R, Festen JM, Plomp R. Effect of temporal envelope smearing on speech reception. *J Acoust Soc Am* 95: 1053–1064, 1994b.

Efron B, Tibshirani RJ. *An Introduction to the Bootstrap*. New York: Chapman and Hall, 1993.

Folstein JR, Van Petten C. Influence of cognitive control and mismatch on the N2 component of the ERP: a review. *Psychophysiology* 45: 152–170, 2008.

Friederici AD, Alter K. Lateralization of auditory language functions: a dynamic dual pathway model. *Brain Lang* 89: 267–276, 2004.

Garofolo J, Lamel L, Fisher W, Fiscus J, Pallett D, Dahlgren N, Zue V. *TIMIT Acoustic-Phonetic Continuous Speech Corpus*. Philadelphia, PA: Linguistic Data Consortium, 1993.

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

Ghitza O, Greenberg S. On the possible role of brain rhythms in speech perception: intelligibility of time-compressed speech with periodic and aperiodic insertions of silence. *Phonetica* 66: 113–126, 2009.

Giraud AL, Poeppel D. Cortical oscillations and speech processing: emerging computational principles and operations. *Nat Neurosci* 15: 511–517, 2012.

Herdman AT, Lins O, Van Roon P, Stapells DR, Scherg M, Picton TW. Intracerebral sources of human auditory steady-state responses. *Brain Topogr* 15: 69–86, 2002.

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

Hillyard SA, Vogel EK, Luck SJ. Sensory gain control (amplification) as a mechanism of selective attention?: electrophysiological and neuroimaging evidence. *Philos Trans R Soc Lond B Biol Sci* 353: 1257–1270, 1998.

Hopf J, Boehler CN, Luck SJ, Tsotsos JK, Heinze H, Schoenfeld MA. Direct neurophysiological evidence for spatial suppression surrounding the focus of attention in vision. *Proc Natl Acad Sci USA* 103: 1053–1058, 2006.

Horton C, D'Zmura M, Srinivasan R. EEG reveals divergent paths for speech envelopes during selective attention. *Int J Bioelectromagnetism* 13: 217–222, 2011.

Howard MF, Poeppel D. Discrimination of speech stimuli based on neuronal response phase patterns depends on acoustics but not comprehension. *J Neurophysiol* 104: 2500–2511, 2010.

Jung TP, Makeig S, Humphries C, Lee TW, McKeown MJ, Iragui V, Sejnowski TJ. Removing electroencephalographic artifacts by blind source separation. *Psychophysiology* 37: 163–178, 2000.

Jung TP, Makeig S, McKeown MJ, Bell AJ, Lee TW, Sejnowski TJ. Imaging brain dynamics using independent component analysis. *Proc IEEE* 89: 1107–1122, 2001.

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. Entrainment of neuronal oscillations as a mechanism of attentional selection. *Science* 320: 110–113, 2008.

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

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

Lalor EC, Power AJ, Reilly RB, Foxe JJ. Resolving precise temporal processing properties of the auditory system using continuous stimuli. *J Neurophysiol* 102: 349–359, 2009.

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.

Lee TW, Girolami M, Sejnowski TJ. Independent component analysis using an extended infomax algorithm for mixed subgaussian and supergaussian sources. *Neural Comput* 11: 417–441, 1999.

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.

McDermott JH. The cocktail party problem. *Curr Biol* 19: R1024–R1027, 2009.

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

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.

Müller N, Schlee W, Hartmann T, Lorenz I, Weisz N. Top-down modulation of the auditory steady-state response in a task-switch paradigm. *Front Hum Neurosci* 3: 1–9, 2009.

Nager W, Kohlmetz C, Joppich G, Möbes J, Münte TF. Tracking of multiple sound sources defined by interaural time differences: brain potential evidence in humans. *Neurosci Lett* 344: 181–184, 2003.

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

Nunez PL, Srinivasan R. *Electric Fields of the Brain: The Neurophysics of EEG*. New York: Oxford Univ. Press, 2006.

Onton J, Makeig S. Information-based modeling of event-related brain dynamics. *Prog Brain Res* 159: 99–120, 2006.

Oostenveld R, Praamstra P. The five percent electrode system for high-resolution EEG and ERP measurements. *Clin Neurophysiol* 112: 713–719, 2001.

Pasley BN, David SV, Mesgarani N, Flinker A, Shamma SA, Crone NE, Knight RT, Chang EF. Reconstructing speech from human auditory cortex. *PLoS Biol* 10: e1001251, 2012.

Peelle JE. The hemispheric lateralization of speech processing depends on what "speech" is: a hierarchical perspective. *Front Hum Neurosci* 6: 1–4, 2012.

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

Picton TW, Hillyard SA, Krausz HI, Galambos R. Human auditory evoked potentials. I. Evaluation of components. *Electroencephalogr Clin Neurophysiol* 36: 179–190, 1974.

Picton TW, Hillyard SA. Human auditory evoked potentials. II. Effects of attention. *Electroencephalogr Clin Neurophysiol* 36: 191–199, 1974.

Picton TW, John MS, Dimitrijevic A, Purcell DW. Human auditory steady-state responses. *Int J Audiol* 42: 177–219, 2003.

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, Foxe JJ, Forde EJ, Reilly RB, Lalor EC. At what time is the cocktail party? A late locus of selective attention to natural speech. *Eur J Neurosci* 35: 1497–1503, 2012.

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

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, Picton TW, Herdman AT, Pantev C. The effect of attention on the auditory steady-state response. *Neurol Clin Neurophysiol* 2004: 22, 2004.

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

Skosnik PD, Krishnan GP, O'Donnell BF. The effect of selective attention on the gamma-band auditory steady-state response. *Neurosci Lett* 420: 223–228, 2007.

Vaughan HG, Ritter W. The sources of auditory evoked responses recorded from the human scalp. *Electroencephalogr Clin Neurophysiol* 28: 360–367, 1970.

Walter VJ, Walter WG. The central effects of rhythmic sensory stimulation. *Electroencephalogr Clin Neurophysiol* 1: 57–86, 1949.

Will U, Berg E. Brain wave synchronization and entrainment to periodic acoustic stimuli. *Neurosci Lett* 424: 55–60, 2007.

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.

Zatorre RJ, Belin P, Penhune VB. Structure and function of auditory cortex: music and speech. *Trends Cogn Sci* 6: 37–46, 2002.