Title

Attention selectively modulates cortical entrainment in different regions of the speech spectrum.

Authors

Baltzell, Lucas
Horton, Cort
Shen, Yi
et al.

Publication Date

2016-08-01

DOI

10.1016/j.brainres.2016.05.029

Peer reviewed

HHS Public Access

Author manuscript

Brain Res. Author manuscript; available in PMC 2017 August 01.

Published in final edited form as:

Brain Res. 2016 August 1; 1644: 203–212. doi:10.1016/j.brainres.2016.05.029.

Attention selectively modulates cortical entrainment in different regions of the speech spectrum

Lucas S. Baltzell[1,*], **Cort Horton**[1], **Yi Shen**[1,3], **Virginia M. Richards**[1], **Michael D'Zmura**[1], and **Ramesh Srinivasan**[1,2]

[1]Department of Cognitive Sciences, University of California, Irvine, California

[2]Department of Biomedical Engineering, University of California, Irvine, California

[3]Department of Speech and Hearing Sciences, Indiana University, Indiana

Abstract

Recent studies have uncovered a neural response that appears to track the envelope of speech, and have shown that this tracking process is mediated by attention. It has been argued that this tracking reflects a process of phase-locking to the fluctuations of stimulus energy, ensuring that this energy arrives during periods of high neuronal excitability. Because all acoustic stimuli are decomposed into spectral channels at the cochlea, and this spectral decomposition is maintained along the ascending auditory pathway and into auditory cortex, we hypothesized that the overall stimulus envelope is not as relevant to cortical processing as the individual frequency channels; attention may be mediating envelope tracking differentially across these spectral channels. To test this we reanalyzed data reported by Horton et al. (2013), where high-density EEG was recorded while adults attended to one of two competing naturalistic speech streams. In order to simulate cochlear filtering, the stimuli were passed through a gammatone filterbank, and temporal envelopes were extracted at each filter output. Following Horton et al. (2013), the attended and unattended envelopes were cross-correlated with the EEG, and local maxima were extracted at three different latency ranges corresponding to distinct peaks in the cross-correlation function (N1, P2, and N2). We found that the ratio between the attended and unattended cross-correlation functions varied across frequency channels in the N1 latency range, consistent with the hypothesis that attention differentially modulates envelope-tracking activity across spectral channels.

Keywords

Entrainment; EEG; Attention; Speech Envelopes

[*]Correspondence to: Social Science Lab (SSL) 184, Univ. of California, Irvine, CA 92697. baltzell@uci.edu.

L.S.B., C.H., Y.S., V.M.R., M.D., and R.S. conceived of and designed research; C.H. performed experiments, L.S.B analyzed data; L.S.B., C.H., Y.S., V.M.R., M.D., and R.S. interpreted results; L.S.B. prepared figures; L.S.B. drafted manuscript; L.S.B., C.H., Y.S., V.M.R., M.D., and R.S. edited, revised, and approved manuscript.

Disclosures: No conflicts of interest, financial or otherwise, are declared by the authors

1. Introduction

Recently, a number of studies have been published examining the effects of attention on neural responses that appear to track the temporal envelope of speech (and non-speech) in the auditory cortex (Kerlin et al., 2010; Ding & Simon, 2012a,b; Mesgarani & Chang, 2012; Ng et al., 2012; Power et. al., 2012; Horton et al., 2013; Horton et al., 2014; Zion-Golumbic et al., 2013; Ding & Simon, 2014; Ding et al., 2014; O'Sullivan, 2014; Di Liberto et al., 2015). This phenomenon is often referred to as cortical "entrainment," and while the underlying mechanisms are still unclear, it is thought to reflect important aspects of temporal processing (for a review, see Ding & Simon, 2014). Pointing to the correspondence between the modulation spectrum of speech and the power spectrum of cortical oscillations, it has been suggested that these oscillations play an active role in parsing the acoustic speech stimulus into discrete syllable-length units for linguistic processing (Ghitza, 2011; Giraud & Poeppel, 2012; Doelling et al., 2014). While this functional claim remains somewhat controversial, it appears clear that the envelope-tracking response reflects attentional mechanisms that enhance the response to the target and suppress the response to the distractor in a complex auditory scene (Ding & Simon, 2012a,b; Mesgarani & Chang, 2012; Horton et al., 2013).

It has been suggested that attention is modulating the envelope-tracking response at the level of the auditory object (e.g. Ding & Simon, 2012a,b; Ding & Simon, 2014). In other words, attention is being applied to some neural reconstruction of the target and distractor auditory objects, formed by integrating information across frequency channels. In support of this position, Ding et al. (2014) showed that degrading the spectro-temporal fine structure (while leaving the temporal envelope intact) led to a reduction in the envelope-tracking response, suggesting that cortical envelope tracking depends on object formation. Additionally, Rimmele et al. (2015) show that degrading the spectro-temporal fine structure also leads to a reduced effect of attention on the envelope-tracking response. However, describing the effect of attention as acting on the neural representation of a formed auditory object may overlook the fact that objects, once formed need to be *maintained* over time. In other words, the object formation process must be continuously updated, and this process would be expected to require the deployment of attention to neural representations prior to their integration into a single object (Winkler et al., 2009).

For instance, before auditory objects can be formed, sounds pass through a bank of peripheral auditory filters, and the resulting spectral (frequency) channels are preserved in the ascending auditory pathway (Kaas et al., 1999; Humphries et al., 2010). These filters can introduce important non-linear transforms, including the introduction of envelopes not contained in the original stimulus (see Ghitza et al., 2013). Furthermore, attention can selectively modulate activity at even the earliest stages of encoding (Maison et al., 2001), and attention can be selectively deployed to particular frequency regions (Mondor & Bregman, 1994). For these reasons, we expect attention to be applied *non*-uniformly across spectral channels as a function of time, consistent with a model of auditory scene analysis that allows for a constant feedback loop between object formation, object selection, and low-level feature representations, with attention being able to influence the object formation process rather than just the object itself (e.g. Winkler et al., 2009).

Furthermore, we might expect that the effect of attention will be more pronounced in spectral channels corresponding to regions of the speech spectrum that are important for intelligibility, rather than simply tracking those regions that contain the most energy. Greenberg et al. (1998) suggest that 1/3-octave spectral channels in the approximately 750-2350 Hz range contribute substantially to speech intelligibility, and while this region is narrower than the speech importance region identified in the ANSI standards (ANSI, 1997), it is clear that spectral regions important for speech perception are not necessarily those that contain the most energy. This is especially true between approximately 1000 and 5000 Hz, where decreases in speech energy are not followed by decreases in speech importance.

The goal of the current study was to examine the extent to which the cortical entrainment reported in Horton et al. (2013) is selective over spectral channels. In the original study, subjects were instructed to listen to one of two speech streams presented from different free-field loudspeakers positioned 45 degrees to the left and right of center. Cortical entrainment was measured for both the attended and unattended speech stimuli as the cross-correlation between the stimulus envelope and the neural response, which recovers a *temporal response function* with distinct peaks corresponding to the typical N1-P2-N2 onset response (Figure 1). Peaks in this temporal response function can be interpreted as delays at which the neural response reliably follows the stimulus envelope. Following the event-related potential literature (Hall III, 2007), we treat these peaks as reflecting distinct neural processes, with later peaks reflecting downstream processes in the cortical auditory hierarchy. As the auditory signal is processed downstream, information is integrated across spectral channels auditory objects are formed (e.g. Rauschecker & Tian, 2000), and since we are examining within-channel processes, we expect find larger effects at earlier latencies (i.e. N1).

In order to decompose the stimulus into spectral channels, we passed the stimulus through a gammatone filterbank with eighteen filters equally spaced on a log scale between 100 to 6246 Hz (Figure 2). At the output of each of these gammatone filters, which are designed to model cochlear filtering, the attended and unattended envelopes were extracted and cross-correlated with the neural response to obtain attended and unattended temporal response functions for each spectral channel.

By focusing on the *ratio* between the attended and unattended envelope-tracking response, we show that the effect of attention on the envelope-tracking response is not uniform across spectral channels in the N1 latency range. This suggests that attention is modulating the envelope-tracking response *within* spectral channels, and is therefore influencing the process of object formation rather than simply applying gain to the object itself. Furthermore, we show significant attentional modulation at high frequencies (1851 – 6246 Hz) where energy is relatively sparse, suggesting that attention is directed to high-importance rather than high-energy regions.

Results

Latency ranges for the N1, P2, and N2 peaks were defined (Figure 3), and a subset of maximally-responding channels were selected to form a region of interest (ROI) within each latency range. Attended and unattended cross-correlation maxima were then selected from

the ROI time series, yielding the functions shown in Figure 4a. A bootstrap was performed to estimate a noise floor for cross-correlation maxima due to chance (see Experimental Procedures). For the purposes of statistical analysis, we collapsed over low, mid, and high frequency regions, shown in Figure 4b. In latency ranges where a significant interaction between attention (attended vs unattended) and frequency region was observed, we quantified the effect of attention as the attended/unattended log-ratio, and performed post-hoc tests on this log-ratio function (Figure 5). We focus on the ratio because this provides a summary effect of attention, reflecting both target (attended stimulus) enhancement and masker (unattended stimulus) suppression, and took the log of the ratio so that the distribution of ratio values were approximately normal.

2.1 The N1 latency range

A 2-factor MANOVA in the N1 latency range revealed a significant interaction between frequency and attention (Pillai's trace = .709, $F[1,9] = 9.7$, $p = .007$, $\eta_p^2 = .709$). Therefore, we examined the simple effect of frequency for both the attended and unattended functions. This post-hoc MANOVA analysis revealed a marginally significant (after Bonferroni correction) simple effect of frequency for the attended function (Pillai's trace = .534, $F[2,8] = 4.59$, $p = .047$, $\eta_p^2 = .534$), and a significant simple effect for the unattended function (Pillai's trace = .815, $F[2,8] = 17.6$, $p = .001$, $\eta_p^2 = .815$).

Having found a significant interaction, we performed paired-comparisons on the attended/unattended log-ratio, which revealed that the envelope-tracking response is significantly smaller in the low frequency region than in the mid ($p = .004$) and high frequency regions, ($p = .005$), but responses are not significantly different between mid and high frequency regions ($p = .71$). These results are shown in Figure 5a.

2.2 The P2 latency range

A 2-factor repeated measures MANOVA in the P2 latency range revealed a significant main effect of attention (Pillai's trace = .702, $F[1,9] = 21.2$ $p < .001$, $\eta_p^2 = .702$), a significant main effect of frequency region (Pillai's trace = .611, $F[2,8] = 6.29$, $p = .023$, $\eta_p^2 = .611$), but no significant interaction between frequency region and attention (Pillai's trace = .211, $F[2,8] = 1.07$, $p = .387$, $\eta_p^2 = .211$). Due to a lack of a significant interaction between frequency and attention, we did not perform post-hoc analyses on the attended/unattended log-ratio. However, we further investigated the significant main effect of frequency region by examining the simple effect of frequency for both the attended and unattended functions. We found that neither the attended (Pillai's trace = .455, $F[2,8] = 3.34$, $p = .088$, $\eta_p^2 = .455$) nor unattended (Pillai's trace = .110, $F[2,8] = .49$, $p = .627$, $\eta_p^2 = .11$) function reached significance.

2.3 The N2 latency range

A 2-factor repeated measures MANOVA in the N2 latency range revealed a significant main effect of attention (Pillai's trace = .495, $F[1,9] = 8.81$ $p = .016$, $\eta_p^2 = .495$), no significant main effect of frequency region (Pillai's trace = .434, $F[2,8] = 3.06$, $p = .103$, $\eta_p^2 = .434$), and a significant interaction between frequency region and attention (Pillai's trace = .668, $F[2,8] = 8.05$, $p = .013$, $\eta_p^2 = .668$). Therefore, we performed paired-comparisons on the

attended/unattended log-ratio but found no significant comparisons (all $p > .05$). These results are shown in Figure 5b. Furthermore, effect sizes were moderate for the low-frequency to high-frequency ($d = .422$) and mid-frequency to high-frequency ($d = .491$) comparisons, suggesting that small effect sizes are not driving the lack of significance.

3. Discussion

The data reported here suggest that the modulatory effects of attention on the neural tracking of speech envelopes can depend on frequency region. The summary effect of attentional modulation was quantified as the ratio between attended and unattended cross-correlation maxima. This ratio showed a significant effect of frequency region in the N1 latency range, but not in the P2 and N2 latency ranges.

3.1 The effect of attention in the N1 latency range

Significant differences in the attended/unattended ratio across frequency regions in the N1 latency range are inconsistent with a model of envelope-tracking that strictly follows the full-band envelope, as such a model predicts that the attended/unattended ratio across spectral channels would remain constant (or flat). This conclusion is further supported by the fact that the simple main effect of frequency region was significant for the unattended function and marginally significant (after Bonferroni correction) for the attended function.

Specifically, mid and high frequency regions show significantly greater attentional modulation than low frequency regions (Figure 5a). If we consider that the stimulus power spectrum (Figure 6a) peaks at mid frequencies, there is an intuitive interpretation of the difference between low and mid frequencies, namely, that attention is deployed in mid frequency channels because these channels contain the most stimulus energy.

If attentional modulation of the envelope-tracking response were simply following stimulus energy however, we would expect to see a difference in the attended/unattended ratio between mid and high frequency regions. The fact that the attended/unattended ratio in the high-frequency region is significantly larger than in the low-frequency region and *not* significantly different than the mid-frequency region suggests that attentional modulation is not following stimulus energy in the high frequency region. If we consider that fricatives provide high-frequency, broadband bursts of energy (Strevens, 1960), and that the cortex is prone to respond to abrupt onsets (Phillips, Hall & Boehnke, 2002), it is perhaps not surprising that attention would be directed to those channels that carry these abrupt onsets, namely, those in the high-frequency region. Recent studies have demonstrated that the timing and frequency content of fricative bursts at an above ∼1500 Hz are crucial for differentiating phonemes (Li, Menon & Allen, 2010; Li et al., 2012). We might also think of these fricative bursts as acoustic landmarks for syllable structure and word boundaries, and therefore particular important in degraded listening environments (Li & Loizou, 2008). Indeed, Doelling et al. (2014) showed that sharp envelope fluctuations drive envelope tracking, and that this tracking correlates with intelligibility.

3.2 Effects of attention in the P2 and N2 latency range

In the P2 latency range, our analysis did not reveal a significant interaction between frequency region and attention, which is to say that the effect of frequency region was not significantly different between the attended and unattended envelope-tracking response. However, in the N2 latency range, a significant interaction was observed, which prompted us to analyze the attended/unattended ratio. Shown in Figure 5b, no pairwise comparisons were significant, which limits our ability to discuss this interaction. While it may be the case that the attended/unattended ratio depends on frequency region, this effect was not robust in our dataset.

Instead, our data suggest that in the N1 latency range, there is a robust difference between the effect of frequency region on the attended and unattended envelop-tracking response, while in the P2 and N2 latency range this difference was not observed. Later latency ranges (P2, N2) reflect downstream processes in the cortical auditory hierarchy (Shahin et al., 2005; Tonnguist-Uhlen, 1996). As the auditory signal is processed downstream, information is integrated across spectral channels (i.e., auditory objects are formed; Rauschecker & Tian, 2000). Therefore, the pattern of attention effects observed here (frequency-specific in N1, non-specific in P2, N2) may reflect a transition from a low-level, tonotopic representation of the signal (N1) to a high-level, object-based representation of the signal (P2, N2).

3.3 Attentional enhancement vs. suppression

We have chosen to focus our discussion thus far on the attended/unattended ratio, as this is a summary effect of attention that can reflect both target (attended stimulus) enhancement and masker (unattended stimulus) suppression. This is motivated in part from a lack of control shape against which to test our attended and unattended envelope-tracking responses across frequency region. However, as shown in Figure 5b, we see that while the attended function rises from low to mid/high frequencies (in the N1 latency range), the unattended function falls. We believe such an effect is consistent with suppression of the competing talker, especially if we consider the attended function as a proxy for a control (Horton et al., 2013). In particular, we might expect greater envelope tracking in the mid-frequency region relative to the low-frequency region, as there is far more energy in the mid-frequency region. If such an assumption is valid, then the fact that the envelope tracking response to the competing talker *decreases* from the low-frequency to mid-frequency region almost certainly reflects attentional suppression.

3.4 Contrast to previous research

There are two results that should be considered relative to the findings reported in the current study. First, Ding & Simon (2012a) failed to find an effect of attention on the shape of the spectral response function, which plots correlation as a function of frequency. In other words, the ratio between the attended and unattended envelope-to-MEG correlations was flat. However, there are a number of differences between our study and theirs. Perhaps most importantly, their data were reported after projecting the data onto a single source, which implicitly filtered the MEG time series. We made no attempt to localize a single source, and our data almost certainly include activity from multiple sources within and outside of auditory cortex (Giard et al., 1994). Furthermore, our cross-correlation analysis

independently computed a temporal response function for each frequency channel, while the spectro-temporal receptive field (STRF) analysis used by Ding & Simon (2012a) fit temporal and spectral response functions with the same model. Second, Mesgarani & Chang (2012), recording ECoG from electrodes on the surface of the superior temporal gyrus (the location of A2), found that the effect of attention was spatially distributed among recording sites, and did not identify any particular regions that were driving the attentional modulation. This means that, to the extent that activity in A2 is tonotopically organized, the effect of attention is distributed rather than localized in frequency. The distribution of the effect of attention however, was not statistically evaluated, and it is therefore difficult to make direct comparisons to our result.

With these results in mind, it is possible that attentional modulation of the envelope-tracking response can occur both within and across spectral channels. Indeed, there is no reason to assume that envelope tracking within spectral channels *precludes* envelope tracking to the full-band (or integrated) envelope.

3.5 Limitations and Suggestions for Future Research

We report that the effect of attention on the neural response to the speech envelope can be frequency dependent, though our analysis only permits a narrow interpretation of this dependency. Because we did not systematically vary the frequency content of our speech stimuli over trials, we cannot suggest that attention is allocated to different frequency bands on a trial-by-trial (or utterance-by-utterance) basis. Instead, our results only suggest that on average, in the N1 latency range, attention modulates the neural envelope-tracking response in a frequency-dependent fashion (Figure 5a). This frequency dependency may reflect a fixed property of the auditory system, or it may represent the average response of utterance-specific attentional modulation. In other words, we don't know whether or not the frequency dependency we observe represents an active tracking of the frequency content of each utterance. Furthermore, our analysis does not rule out the possibility that attention is modulating the neural response to the full-band envelope of the integrated auditory object in addition to modulating the envelope-tracking response within individual frequency channels. Indeed, as explained above, there is no reason to suspect that attentional modulation of the envelope-tracking response may occur *within* and *across* spectral channels. The first of these limitations can be addressed with a follow up study that systematically fixes the spectra of the attended and unattended speech stimuli across trials, and while the second limitation may prove difficult to address with EEG, techniques with better spatial resolution may be able to resolve this issue.

3.6 Conclusions

In a multi-talker listening environment, the envelope-tracking response to the attended talker is larger than the response to the unattended talker in three latency ranges corresponding to the N1, P2, and N2 peaks in the auditory evoked response. Crucially, in the N1 latency range, attention differentially modulates the envelope-tracking response in different frequency regions, suggesting that attention is deployed differentially across spectral channels. This result is inconsistent with the suggestion that attention is deployed

exclusively to the envelope of an integrated auditory object, and instead suggests that attention influences the process of object formation.

4. Experimental Procedures

The goal of the current study is to examine the extent to which the cortical entrainment reported in Horton et al. (2013) is selective in frequency. The following provides a brief description of the methods reported in Horton et al. (2013), and a detailed description of those methods novel to this study.

4.1 Participants

All experimental procedures were approved by the Institutional Review Board of the University of California, Irvine. Ten young adults (2 female; age: 21-29) participated in the study, although one had to be excluded due to excessive EEG artifacts.

4.2 Task and Stimuli

Each participant sat in a sound-attenuated testing chamber and faced a computer monitor that was flanked on either side by a loudspeaker. At the start of each trial, the subject was presented with a visual cue to attend to either the left or right speaker (chosen at random) while maintaining visual fixation on a cross in the center of the monitor. On each trial, two independent series of spoken sentences were played from two loudspeakers separated at a 90 degree angle. To build these speech stimuli, sentences were drawn at random from the TIMIT speech corpus (Garofolo et al., 1993) and concatenated until the total length of each speech stimulus exceeded 22 seconds. At the end of each trial, subjects were shown the transcript of a sentence from the trial, and were asked to indicate via a button press whether the sentence was played on the attended side. Subjects completed 320 trials (8 blocks, 40 trials per block), with the exception of one subject who only completed 240 trials due to equipment failure.

4.3 EEG Recording and Pre-Processing

High-density EEG (128 channels) was recorded with equipment from Advanced Neuro Technology. Electrodes were placed following the international 10/5 system (Oostenveld & Praamstra, 2001), and all channel impedances were kept below 10 kΩ. The EEG data was average-referenced and filtered offline with a passband of 2 to 40 Hz. The filtered data were then down-sampled from 1024 Hz to 256 Hz and segmented into individual trials which were 20 seconds long, beginning one second after the onset of the sentences. This delay was incorporated to remove any effect of a synchronous onset between the left and right speech stimuli.

4.4 Gammatone Filtering and Envelope Extraction

In order to simulate frequency selectivity of the auditory system, each speech stimulus was passed through a gammatone filterbank (Slaney, 1993), a well-established model of peripheral auditory filtering (for a recent review, see Lyon et al., 2010). Shown in Figure 2, center frequencies of the filters were equal-log-spaced from 100 to 6246 Hz (18 total filters). To extract the envelope, the output of each filter was then Hilbert transformed, and its

magnitude was low-pass filtered (30 Hz). The resulting envelope was then high-pass filtered at 2 Hz to remove the DC component. Artifacts were removed using the Infomax ICA algorithm from the EEGLAB toolbox (Delorme and Makeig, 2004). The speech stimuli from the TIMIT database were originally sampled at 16000 Hz, so center frequencies close to the Nyquist rate (8000 Hz) were not considered.

4.5 Cross-correlation analysis

The neural response to speech stimuli was quantified by computing the cross-correlation functions between the EEG and the envelopes of the attended and unattended speech stimuli in each gammatone-filtered frequency band (see Ahissar et al., 2001 & Power et al, 2012). The cross-correlation function measures the similarity between two discrete signals f and g over a range of delays n.

$$(f \star g)(n) = \sum_{m=-\infty}^{\infty} \frac{f[t]g[n+t]}{\mathrm{std}(f)\mathrm{std}(g)}.$$

Since the cross-correlation is normalized between 0 and 1, the absolute magnitudes of f and g are not reflected in ($f \star g$). The cross-correlation functions between the EEG and the stimulus envelope strongly resemble the N1-P2-N2 response of a typical auditory evoked potential (AEP), which is consistent with our expectation that the AEP reflects the basic response characteristics of the auditory system (Figure 1). For every trial, for each subject, recordings from each channel of the EEG was cross-correlated with both the attended and the unattended stimulus envelopes, and cross-correlation values were Fisher z-transformed to approximate a normal distribution, following the analysis in Horton et al. (2013).

Cross-correlation functions were then averaged across trials, and maximum values were extracted in each of three latency ranges that corresponded to cross-correlation peaks that resembled a typical AEP. Thirty-two out of 128 channels with the largest attended cross-correlation values were identified from grand-averaged subject data separately for three latency ranges corresponding to peaks in the AEP (labeled N1, P2, N2). A large number of channels (32) were included so that broad activity on both sides of the dipoles, shown in Figure 3, could be captured. Having chosen these 32 channels, for each subject, the mean of the absolute value of the attended and unattended cross-correlation functions were computed in each latency range (90 ± 25 ms, 200 ± 25 ms, and 350 ± 25 ms). Taking the absolute value allowed us to average across channels without respect to polarity. From this "composite" channel, the maximum value was selected for both listening conditions (attended, unattended) in all three latency ranges for each of the 18 gammatone-filtered envelopes and the unfiltered envelope.

To estimate a noise floor for these maxima, a bootstrap simulation was performed. A control distribution was constructed by replacing the attended and unattended stimuli on each trial with random stimuli not presented on that trial, and performing the same analysis just described over 1000 iterations. 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-correlations were due purely to chance. Maximum values were considered significantly non-zero if they fell outside the 99.5^{th} or 0.5^{th} percentiles of this distribution.

4.6 Stimulus Properties

4.6.1 Stimulus Power Spectrum—Average power at the output of each gammatone filter (before envelope extraction) is shown in Figure 4c. We see that power increases sharply on a log scale from low frequencies to a peak at around 600 Hz (mid frequency), and decreases at high frequencies. While this pattern is roughly quadratic, we want to point out the asymmetry of the low and high-frequency tails. Specifically, the lowest frequency filters have nearly half the power as highest frequency filters.

4.6.2 Stimulus-to-Stimulus Correlations—The main goal of this analysis was to explore the extent to which envelopes extracted from peripheral channels are tracked in the cortex in a selective attention task. However, since the envelopes at the output of each filter are not uniformly correlated with the full stimulus envelope, any effect of center frequency on the strength of correlation might merely reflect the extent to which cortical tracking of the full stimulus envelope is correlated with the envelopes at different center frequencies. Shown in Figure 6c, correlations were highest between ∼500-700 Hz, dropping off at higher and lower center frequencies. Note that, due to overlap between adjacent gammatone filters (Figure 2), neighboring envelopes tend to be correlated with one another (Figure 6b).

If the envelope-tracking response we observe is indeed a tracking of the full-band envelope, we expect that the shape of *both* the attended an unattended cross-correlation-maximum-by-frequency functions follow the stimulus-to-stimulus-correlation function in Figure 6c. As we have quantified the effect of attention in our analysis as the log-ratio between the attended and unattended cross-correlation maxima, we may restate this expectation as a prediction that the attended/unattended log-ratio-by-frequency function will be flat.

4.7 Filtered Cross-Correlation Functions

Figure 7 shows the cross-correlation functions for the attended, unattended, and control stimuli in four individual frequency channels (center frequencies range from 207 – 4000 Hz). As in Figure 1, which showed the cross-correlation functions for the attended, unattended, and control *full-band* stimuli, distinct peaks in the temporal structure can be observed in the both the attended and unattended cross-correlation functions for each frequency channel.

4.8 Statistical Procedure

Our choice of statistical procedure was motivated by two concerns. First, our independent measures had a covariance structure that was not compound symmetric (Figure 6b). Second, the number of independent measures (i.e. gammatone filters, see Figure 2) was selected somewhat arbitrarily, albeit with the goal of maximizing coverage of the spectrum while minimizing filter overlap. In other words, our decomposing of the speech stimulus into eighteen different spectral bands was simply a convenience, and runs the risk of artificially inflating the number of independent measures we use in our analysis.

We first considered using a linear mixed-effects model that allowed us to specify an autoregressive covariance structure for the fixed effect of center frequency. However, this model, like the ANOVA, adjusts degrees of freedom based on number of independent measures, so even if we accounted for the covariance structure, such an approach runs the risk of artificially inflating degrees of freedom and thus artificially inflating significance.

We decided it was more reasonable to reduce the data into three independent spectral channels instead of eighteen, as this would alleviate both concerns. First, by collapsing (averaging) over low, mid, and high frequencies (on a log scale), we restrict the covariance problem to two borders of these three frequency regions. Second, we reduce the number of independent measures down to three, which is far more conservative in terms of degrees of freedom, and allows us to run a standard multivariate ANOVA (Vasey & Thayer, 1987).

Furthermore, a collapse over low, mid, and high frequencies follows naturally from three aspects of the natural speech. First, the power spectrum of the speech (e.g. Figure 4c) has relatively little energy at low and high frequencies, with most of the energy in the mid frequencies, and if we expect the envelope-tracking response to follow stimulus energy, such a division is appealing. Second, fundamental frequencies for adult male and female talkers do not typically exceed 300 Hz, and first formants do not typically fall below 400 Hz, establishing a natural point of division between low and mid frequencies (Titze, 1994). Third, while there is substantial overlap between frequency regions important for the perception of vowels and consonants, spectral information in the 400-1500 Hz range is crucial for the perception of vowels, and bursts of frication in the 1500 Hz and above range are crucial for consonant identification (see Li, Menon & Allen, 2010), again forming a somewhat natural division between mid and high frequencies. Therefore, we collapsed across the lower six (100 – 338 Hz), the middle six (430 – 1452 Hz), and the highest gammatone filters (1851 – 6246 Hz), effectively reducing the data from eighteen independent frequency regions down to three (low, mid and high).

Using these frequency ranges (spectral channels), we ran a 2-factor (low/mid/high × attended/unattended) multivariate ANOVA. We used a multivariate approach because this allows us fit the covariance structure empirically, rather than assuming compound symmetry (Vasey & Thayer, 1987). This analysis was run separately for the N1, P2 and N2 latency ranges. In ranges that revealed a significant interaction between frequency and attention, post-hoc analyses were performed on the attended/unattended log-ratio (Figure 5). We decided to characterize the effect of attention as an attended/unattended ratio because it effectively removes any effect of the envelope-tracking response not due to attention. In other words, both attentional enhancement of the target and attentional suppression of the masker will be reflected in the ratio, and it is the relative strength of the attended and unattended envelope-tracking response in each frequency region that best summarizes the effect of attention. The log-transform was applied so that the distribution of ratios were approximately normal.

Acknowledgments

This work was supported by National Institute of Mental Health Grant 2R01-MH-68004, Army Research Office Grant ARO 54228-LS-MUR, and National Institute of Health NIDCD R21 DC013406. We would like to thank Jon Venezia for helpful statistical advice.

References

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

ANSI S3.5. Methods for the calculation of the speech intelligibility index. American National Standards Institute. 1997

Delorme A, Makeig S. EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. J Neurosci Methods. 2004; 134:9–21. [PubMed: 15102499]

Di Liberto GM, O'Sullivan J, Lalor E. Low-frequency cortical entrainment to speech reflects phoneme-level processing. Cur Biol. 2015; 25:2457–2465.

Ding N, Simon JZ. Neural coding of continuous speech in auditory cortex during monaural and dichotic listening. J Neurophysiol. 2012a; 107:78–89. [PubMed: 21975452]

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

Ding N, Simon JZ. Cortical entrainment to continuous speech: functional roles and interpretations. Front Hum Neurosci. 2014; 8(311):1–7. [PubMed: 24474914]

Doelling KB, Arnal LH, Ghitza O, Poeppel D. Acoustic landmarks drive delta–theta oscillations to enable speech comprehension by facilitating perceptual parsing. NeuroImage. 2014; 85:761–768. [PubMed: 23791839]

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. 2011; 2:130. [PubMed: 21743809]

Ghitza O, Giraud AL, Poeppel D. Neuronal oscillations and speech perception: critical-band temporal envelopes are the essence. Front Hum Neurosci. 2013; 6:340. [PubMed: 23316150]

Giard MH, Perrin F, Echallier JF, Thevenet M, Froment JC, Pernier J. Dissociation of temporal and frontal components in the human auditory N1 wave: A scalp current density and dipole model analysis. Electrophysiology and Clinical Neurophysiology. 1994; 92:238–252.

Greenberg, S.; Arai, T.; Silipo, R. Speech intelligibility derived from exceedingly sparse spectral information; Proceedings of the fifth international conference on spoken language processing; 1998. p. 74-77.

Hall, JW, III. New Handbook of Auditory Evoked Responses. Boston, MA: Pearson Education, Inc; 2007.

Horton C, D'Zmura M, Srinivasan R. Suppression of competing speech through entrainment of cortical oscillations. J Neurophysiol. 2013; 109:3082–3093. [PubMed: 23515789]

Horton C, Srinivasan R, D'Zmura M. Envelope responses in single-trial EEG indicate attended speaker in a “cocktail party”. Journal of Neural Engineering. 2014; 11:1–22.

Humphries C, Liebenthal E, Binder JR. Tonotopic organization of human auditory cortex. NeuroImage. 2010; 50:1202–1211. [PubMed: 20096790]

Kaas JH, Hackett TA, Tramo MJ. Auditory processing in primate cerebral cortex. Current Opinion in Neurobiology. 1999; 9:164–170. [PubMed: 10322185]

Kerlin JR, Shahin AJ, Miller LM. Attentional gain control of ongoing cortical speech representations in a “cocktail party”. J Neurosci. 2010; 30:620–628. [PubMed: 20071526]

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. Euro J Neurosci. 2012; 35:1487–1503.

Li N, Loizou PC. The contribution of obstruent consonants and acoustic landmarks to speech recognition in noise. J Acoust Soc Am. 2008; 124:3947–3958. 2008. [PubMed: 19206819]

Li F, Menon A, Allen JB. A psychoacoustic method to find the perceptual cues of stop consonants in natural speech. J Acoust Soc Am. 2010; 127:2599–2610. [PubMed: 20370041]

Li F, Trevino A, Menon A, Allen JB. A psychoacoustic method for studying the necessary and sufficient perceptual cues of American English fricative consonants in noise. J Acoust Soc Am. 2012; 132:2663–2675. [PubMed: 23039459]

Lyon, RF.; Katsiamis, AG.; Drakakis, EM. History and future of auditory filter models; IEEE International Conference on Circuits and Systems; 2010. p. 3809-3812.

Mesgarani N, Chang EF. Selective cortical representation of attended speaker in multi-talker speech perception. Nature. 2012; 485:233–236. [PubMed: 22522927]

Mondor TA, Bregman AS. Allocating attention to frequency regions. Perception & Psychophysics. 1994; 56:268–276. [PubMed: 7971127]

Ng BSW, Schroeder T, Kayser C. A precluding but not ensuring role of entrained low-frequency oscillations for auditory perception. J Neurosci. 2012; 32:12268–12276. [PubMed: 22933808]

Oostenveld R, Praamstra P. The five percent electrode system for high-resolution EEG and ERP measurements. Clin Neurophysiol. 2001; 112:713–719. [PubMed: 11275545]

O'Sullivan JA, Power AJ, Mesgarani N, Rajaram S, Foxe JJ, Shinn-Cunningham BG, et al. Lalor EC. Attentional selection in a cocktail party environment can be decoded from single-trial EEG. Cerebral Cortex. 2014; 1:10.

Patterson RD. Auditory filter shapes derived with noise stimuli. J Acoust Soc Am. 1976; 59:640–654. [PubMed: 1254791]

Peelle JE, Gross J, Davis MH. Phase-locked responses to speech in human auditory cortex are enhanced during comprehension. Cereb Cortex. 2013; 23:1378–1387. [PubMed: 22610394]

Phillips DP, Hall SE, Boehnke SE. Central auditory onset responses, and temporal asymmetries in auditory perception. Hear Res. 2002; 167:192–205. [PubMed: 12117542]

Rauschecker JP, Tian B. Mechanisms and streams for processing of “what” and “where” in auditory cortex. P Natl Acad Sci USA. 2000; 97:11800–11806.

Rimmele JM, Zion-Golumbic EZ, Schröger E, Poeppel D. The effects of selective attention and speech acoustics on neural speech-tracking in a multi-talker scene. Cortex. 2015; 68:144–154. [PubMed: 25650107]

Shahin A, Roberts LE, Pantev C, Trainor LJ, Ross B. Modulation of P2 auditory-evoked responses by the spectral complexity of musical sounds. Neuroreport. 2005; 16:1781–1785. [PubMed: 16237326]

Slaney M. An efficient implementation of the Patterson-Holdsworth auditory filter bank. Apple Computer, Perception Group, Tech Rep. 1993; 35

Strevens P. Spectra of fricative noise in human speech. Language and Speech. 1960; 3:32–49.

Titze, IR. Principles of Voice Production. Englewood Cliffs, NJ: Prentice Hall; 1994.

Tonnquist-Ulen I. Topography of auditory evoked long-latency potentials in children with severe language impairment: The P2 and N2 components. Ear & Hearing. 1996; 17:314–326. [PubMed: 8862969]

Vasey MW, Thayer JF. The continuing problem of false positives in repeated measures ANOVA in psychophysiology: A multivariate solution. Psychophysiology. 1987; 24:479–486. [PubMed: 3615759]

Winkler I, Denham SL, Nelken I. Modeling the auditory scene: Predictive regularity representations and perceptual objects. Trends in Cognitive Sciences. 2009; 13:532–540. [PubMed: 19828357]

Zion-Golumbic EM, Ding N, Bickel S, Lakatos P, Schevon CA, McKhann GM, Goodman RR, Emerson R, Mehta AD, Simon JZ, et al. Mechanisms underlying selective neuronal tracking of attended speech at a “cocktail party”. Neuron. 2013; 77:980–991. [PubMed: 23473326]

Highlights

- Attention modulates the envelope-tracking response *within* spectral channels
- We show that this effect is limited to the N1 latency range
- Attention tracks speech-importance rather than stimulus energy

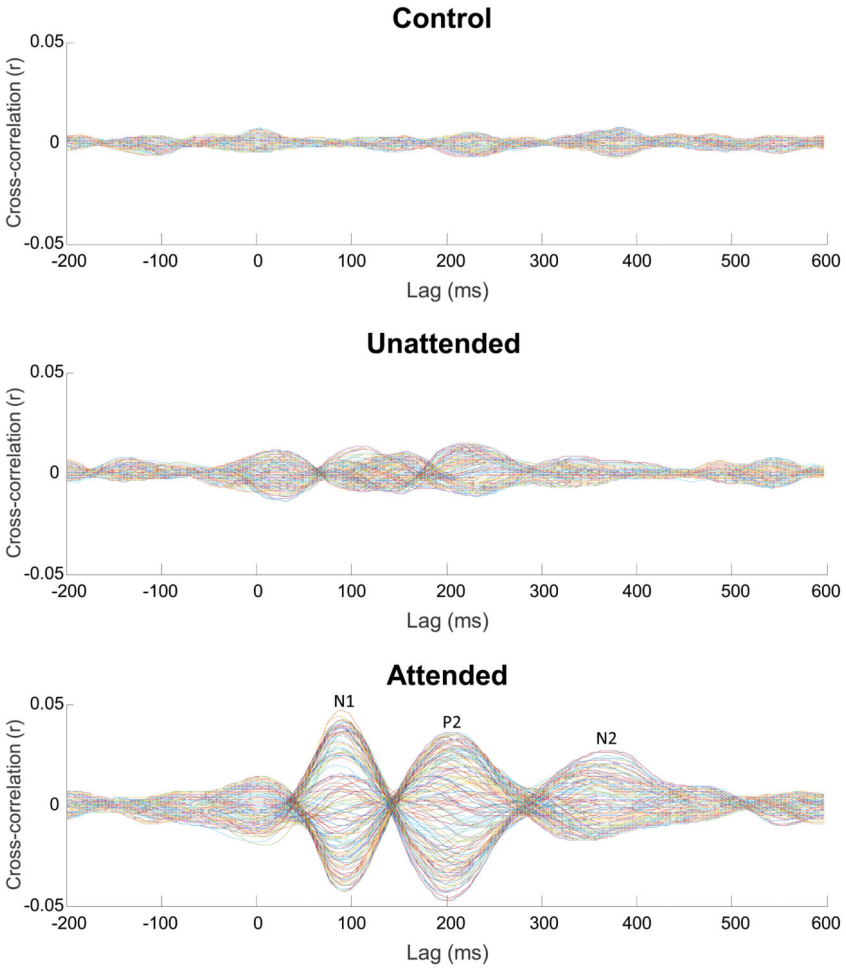

Figure 1.

Cross-correlations between original speech envelopes and EEG activity at 128 recording channels. While we computed cross-correlations with delays from -1000 to +1000 ms, we show only -200 to +600 ms for viewing convenience, and because no significant peaks exist outside of this range. These cross-correlations generate temporal response functions that recover the N1-P2-N2 auditory evoked response, and while the response is clearest in the attended temporal response function, this pattern can also be observed in the unattended temporal response function.

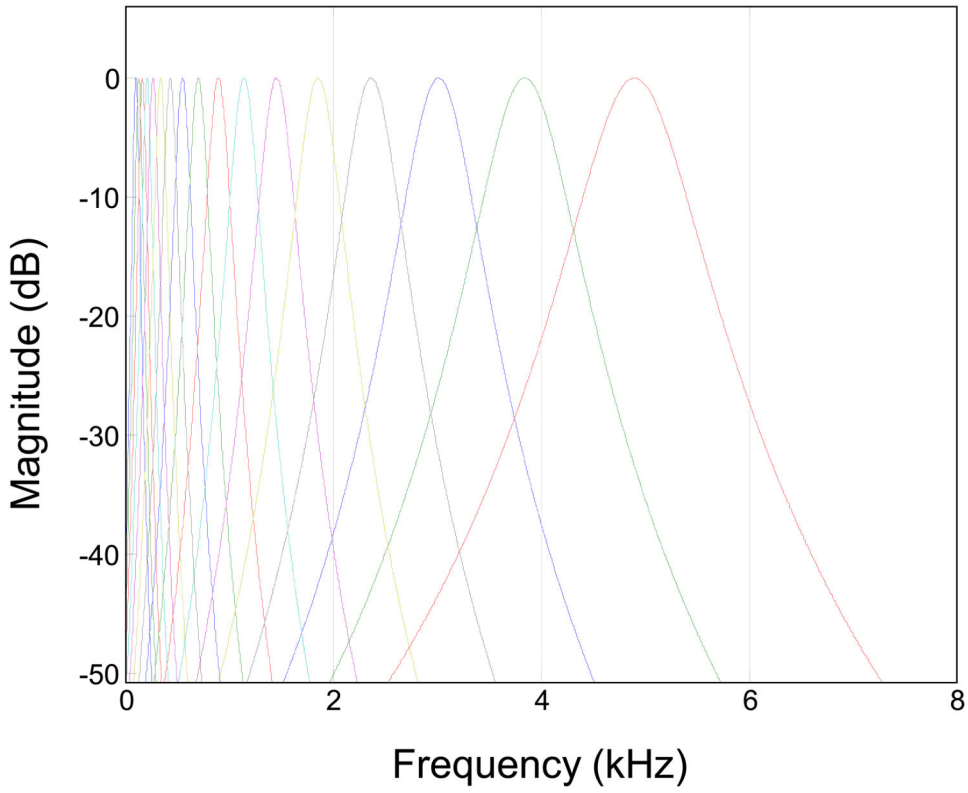

Figure 2.

The frequency response functions of the gammatone filters used in the experiment. On a linear frequency axis, bandwidths increase with increasing center frequency. The matlab code used to generate the gammatone filter coefficients was derived from Slaney (1993).

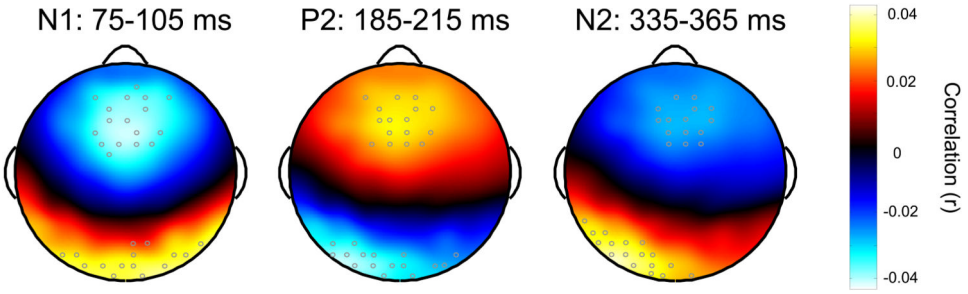

Figure 3.

Scalp topographies for cross-correlation values at all recording sites averaged over latency ranges corresponding to N1, P2, and N2. For each range, a clear anterior-posterior dipole is observed.

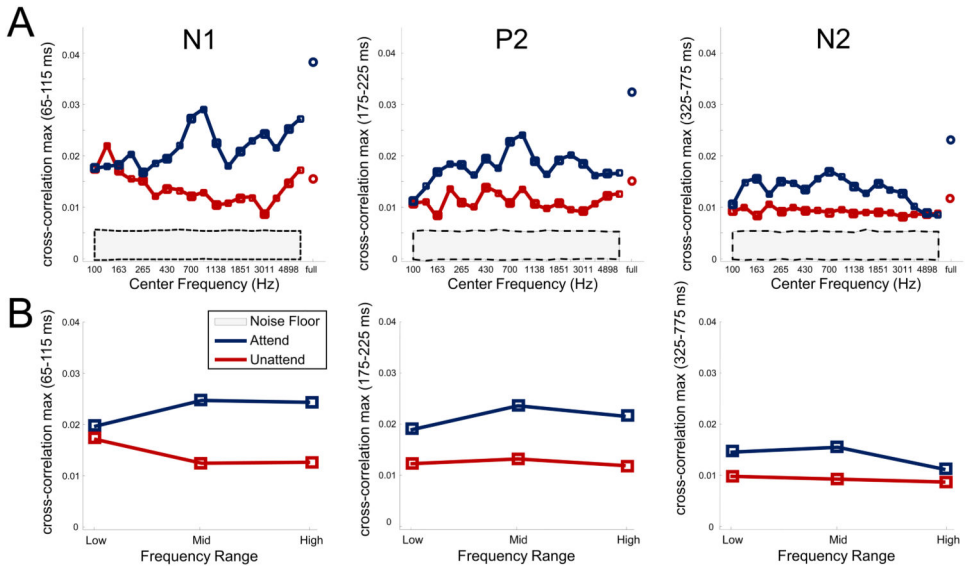

Figure 4.

[**A**] Cross-correlation maxima as a function of gammatone filter center frequency for latencies corresponding to the N1 (90 ± 25 ms), P2 (200 ± 25 ms) and N2 (350 ± 25 ms) peaks. The noise floor (gray) shows the range of correlation values that would occur by chance if the stimulus envelope is unrelated to the EEG. [**B**] The data-reduced version of [A], collapsed into low (100 – 338 Hz), mid (430 – 1452 Hz), and high (1851 – 6246 Hz) frequency regions.

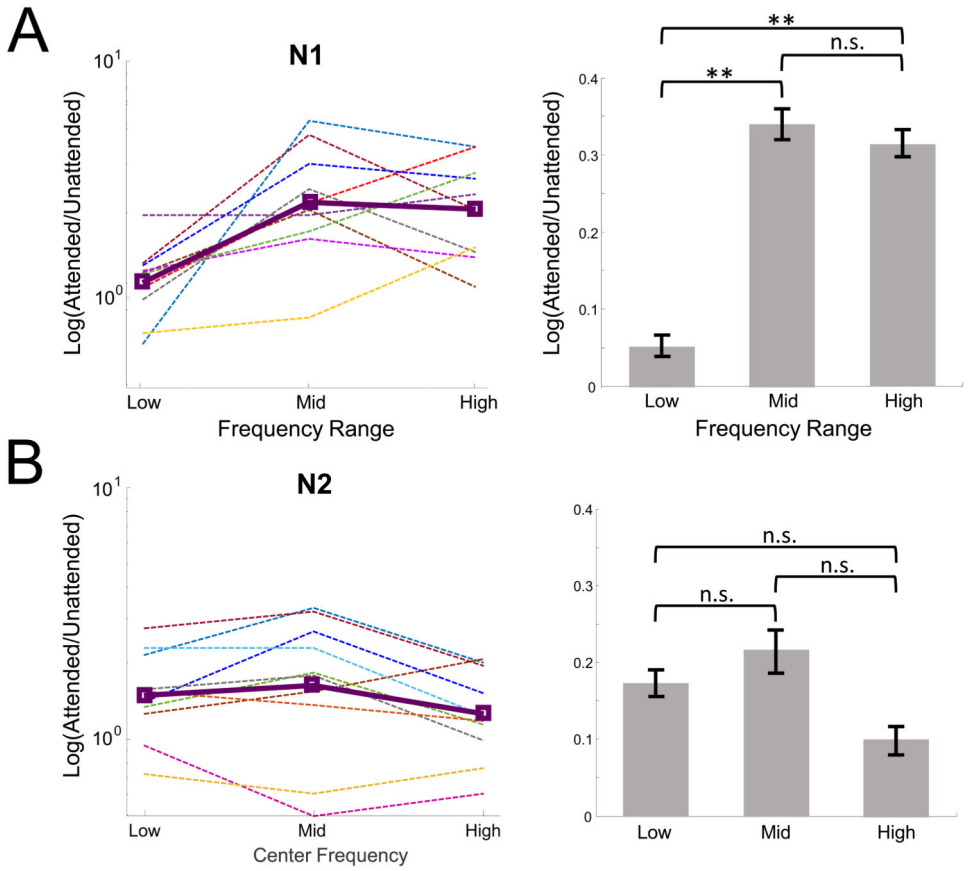

Figure 5.

[**A**] Log-ratios between attended and unattended cross-correlation maxima as a function of frequency region (Low: 100–338 Hz; Mid: 430–1452 Hz; High: 1851–6246 Hz) in the N1 latency range. The solid line indicates the grand average, and each individual dotted line represents an individual subject. On the right of this plot is a bar graph showing the outcome of paired-comparison post-hoc tests. Error bars represent standard errors of them mean. [**B**] Same as [**A**] but for the N2 latency range.

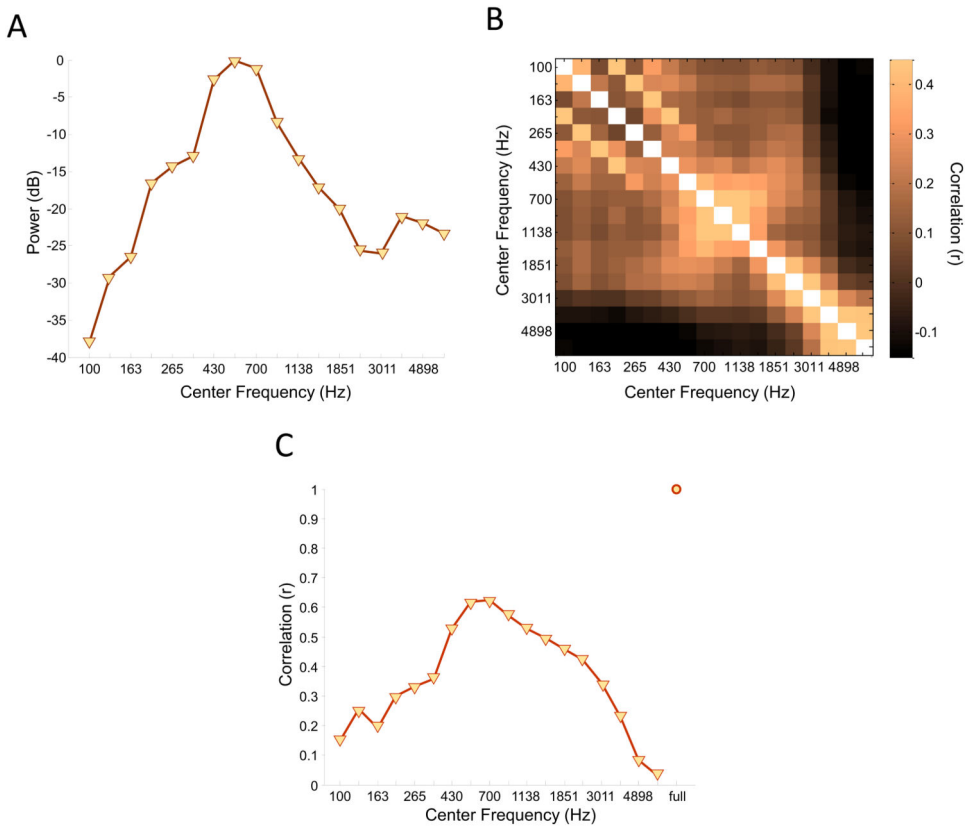

Figure 6.

[A] Total power in each gammatone filter. [**B**] Correlation (normalized covariance) matrix for the same data. [**C**] Correlations between the envelope at the output of each gammatone filter with the original (full-band) stimulus envelope. This stimulus-to-stimulus correlation function can be thought of as the shape of the expected stimulus envelope-to-EEG cross-correlation by frequency function if the full-band stimulus envelope were being entrained.

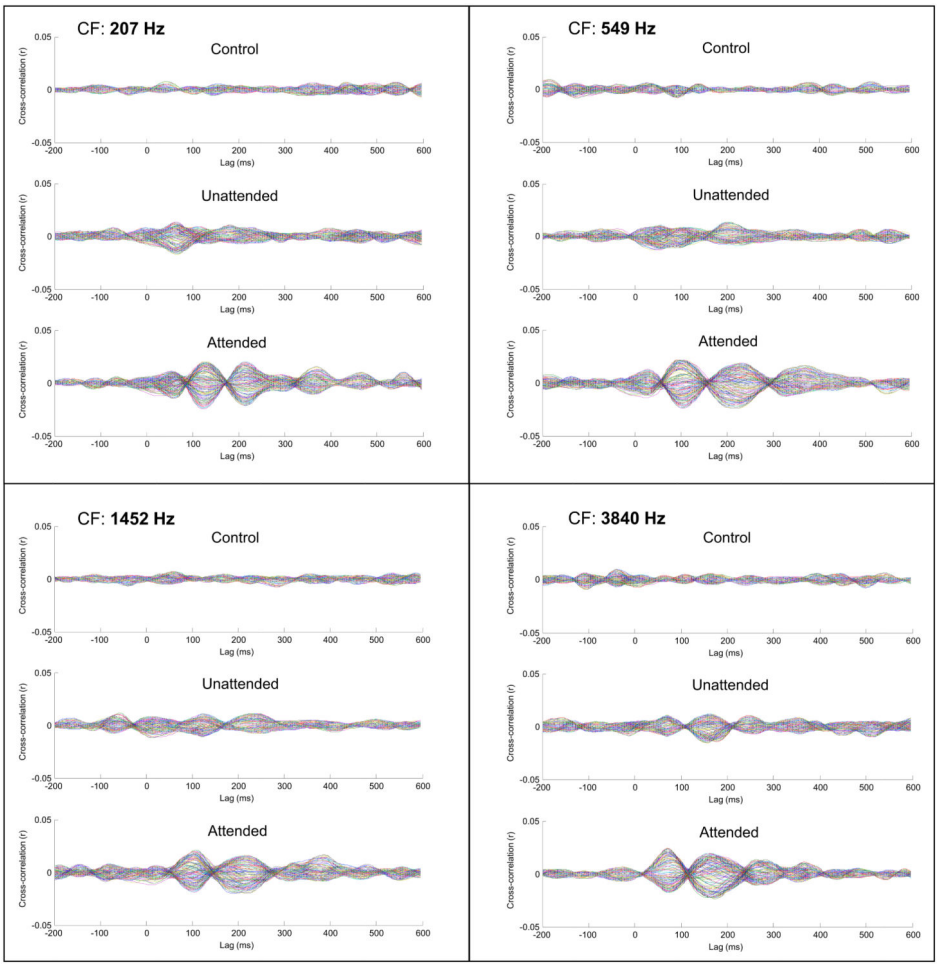

Figure 7.

Cross-correlation functions between original speech envelopes and EEG activity at 128 recording channels for four representative frequency channels (CF) that span the range of CFs included in our analysis. Notice that both the attended and unattended cross-correlation functions show significant structure in the ~65-365 ms latency range, while the control cross-correlation functions do not.