Published in final edited form as:
Hear Res. 2014 October ; 0: 73–81. doi:10.1016/j.heares.2014.07.009.

Differential modulation of auditory responses to attended and unattended speech in different listening conditions

Ying-Yee Kong, PhD[a,b], **Ala Mullangi**[b], and **Nai Ding**[c]

[a]Department of Speech Language Pathology & Audiology, Northeastern University, Boston, MA 02115, United States

[b]Bioengineering Program, Northeastern University, Boston, MA 02115, United States

[c]Department of Psychology, New York University, NY 10012, United States

Abstract

This study investigates how top-down attention modulates neural tracking of the speech envelope in different listening conditions. In the quiet conditions, a single speech stream was presented and the subjects paid attention to the speech stream (active listening) or watched a silent movie instead (passive listening). In the competing speaker (CS) conditions, two speakers of opposite genders were presented diotically. Ongoing electroencephalographic (EEG) responses were measured in each condition and cross-correlated with the speech envelope of each speaker at different time lags. In quiet, active and passive listening resulted in similar neural responses to the speech envelope. In the CS conditions, however, the shape of the cross-correlation function was remarkably different between the attended and unattended speech. The cross-correlation with the attended speech showed stronger N1 and P2 responses but a weaker P1 response compared with the cross-correlation with the unattended speech. Furthermore, the N1 response to the attended speech in the CS condition was enhanced and delayed compared with the active listening condition in quiet, while the P2 response to the unattended speaker in the CS condition was attenuated compared with the passive listening in quiet. Taken together, these results demonstrate that top-down attention differentially modulates envelope-tracking neural activity at different time lags and suggest that top-down attention can both enhance the neural responses to the attended sound stream and suppress the responses to the unattended sound stream.

Keywords

auditory attention; cortical entrainment; suppression; sound segregation; event related potentials

Corresponding Author: Ying-Yee Kong, Ph.D., Department of Speech Language Pathology & Audiology, 226 Forsyth Building, 360 Huntington Ave., Northeastern University, Boston, MA 02115, United States, Tel: +1 (617) 373-3704, Fax: +1 (617) 373-2239, yykong@neu.edu.

1. Introduction

Top-down attention plays an important role in auditory perception in complex listening environments. Using an event-related design, previous electroencephalography (EEG) and magnetoencephalography (MEG) studies showed greater brain responses to attended sounds relative to the responses to unattended sounds (e.g., Hillyard et al., 1973; Näätänen, 1992; Alain & Woods, 1994; Teder-Sälejärvi et al., 1999; Snyder et al., 2006). The attentional effect can occur at various processing stages. For example, the effects of attentional modulation appear as early as ~100 msec in some experiments (e.g., Hillyard et al., 1973; Hansen & Hillyard, 1988; Alho et al., 1994; Melara et al., 2002; Choi et al., 2014) but later in other experiments (>150 msec; e.g, Picton & Hillyard, 1974; Neelon et al., 2006; Snyder et al., 2006; Ross et al., 2010). Stimulus properties and the subject's task are likely to determine the latency of top-down attentional modulations. In addition to the amplification of brain responses to the attended signals, responses to the unattended signals are sometimes attenuated, suggesting active suppression mechanisms for irrelevant stimuli (e.g., Rif et al., 1991; Alain et al., 1993; Alho et al., 1994; Alain & Woods, 1994; Bidet-Caulet et al., 2010). The effect of suppression is usually found in a later processing stage about 200 msec post stimulus onset.

Recent works have extended neurophysiology studies on top-down auditory attention from using relatively simple sounds, e.g. tones, to using more ecologically realistic stimuli, such as continuous speech (see a recent review by Ding & Simon, 2014). When a single speech stream is presented in a quiet listening environment, the neural responses from the auditory cortex phase lock to the temporal envelope of the speech signal (e.g., Luo & Poeppel, 2007; Aiken & Picton, 2008; Howard & Poeppel, 2010; Lalor & Foxe, 2010; Pasley et al., 2012). When two speech streams are presented simultaneously, neural activity shows stronger phase locking to the temporal envelope of the attended speech stream, compared with the phase locking to the unattended speech (Kerlin et al., 2010; Ding & Simon, 2012a, 2012b; Mesgarani & Chang, 2012; Horton et al., 2013; O'Sullivan et al., 2014).

Although most previous studies have shown attentional modulation of speech tracking neural activity (see, however, Zion Golumbic et al., 2013), the latency of the attentional modulation effects has been controversial. While some studies reported an early attentional gain effect at a time lag around 100 msec (Ding & Simon, 2012a, 2012b), others reported a longer-latency attentional effect near 200-msec time lag (Power et al., 2012; O'Sullivan et al., 2014). Furthermore, more complicated patterns were observed by Horton et al. (2013), who reported that the EEG responses were correlated with the attended and unattended speech streams with opposite signs at time lags between 150 and 450 msec. Based on the neurophysiological findings that the phase of slow neural oscillations modulates the excitability of neurons (Lakatos et al., 2008, 2013; Schroeder & Lakatos, 2009), Horton et al. (2013) suggested that the opposite polarities of the cross-correlations reflect enhancement of the attended speech and suppression of the unattended speech.

The goal of the present study is to investigate whether top-down attention differentially modulates neural tracking of the speech envelope when the target speech stream is competed with different types of sensory interference. Specifically, when the properties of the sensory

interference varied, we tested if the effect of top-down attention may change from modulating the response gain (e.g., Ding & Simon, 2012b) to modulating the response timing (Horton et al., 2013; O'Sullivan et al., 2014), or to not significantly modulating cortical activity at all (Zion Golumbic et al., 2013).

First, we investigated whether cortical responses may be differentially modulated by attention when competing information was presented via the same or different sensory modalities. In the cross-modality condition, a narrated story was presented in a quiet listening environment, and the subjects were instructed to either listen to the story or watch a silent movie instead. In the within-modality condition, the subjects heard a mixture of two simultaneous speakers, one male and one female, and had to selectively attend to one of them based on the instruction. Second, we compared the neural responses to an attended or unattended speech stream when that speech stream was presented together with a speech stream of the same sound intensity, a speech stream of a lower sound intensity, or in a quiet listening environment. Using these conditions, we probed how irrelevant information sources are filtered out by top-down attention when competitors are presented from different sensory modalities, and how the filtering process may depend on the amount of interference within the auditory modality.

2. Methods

2.1. Subjects

Eight normal-hearing, right-handed (Oldfield, 1971), adult native speakers of American English between the ages of 21 and 36 years participated in the study. This study was conducted according to the protocols approved by the Institutional Review Board of Northeastern University. Written informed consent was obtained prior to the experiment.

2.2. Test stimuli and Procedures

Auditory stimuli were continuous speech extracted from two chapters in a public domain children's book, "A Child's History of England" by Charles Dickens (http://librivox.org/a-childs-history-of-england-by-charles-dickens/), narrated by one male and one female speaker. The first chapter was 22, read by a male speaker, and the second was chapter 35, read by a female speaker. The sampling rate of the recordings was 22.05 kHz. All silent intervals longer than 300 msec were shortened to 300 msec to maintain continuous flow of the speech streams. The extracted passages were divided into sections with durations of approximately one minute each. The actual length of the segment varied slightly to include a complete sentence. All of the 1-min speech segments were normalized to have equal root-mean-square (RMS) amplitude. In addition, for the competing speaker conditions described below, speech mixtures were constructed by mixing two speakers digitally with one speaker beginning 1 sec after the other speaker, and both speakers ending the same time. The RMS level of one speaker was fixed in the mixture, while the other speaker (the speaker with a delayed start) was either the same or 6 dB weaker, resulting in two target-to-masker ratio (TMR) conditions. All stimuli were presented diotically using insert earphones. All experiments were conducted in a double-walled sound-isolated booth.

There were two main experiments – speech comprehension in quiet and speech comprehension in a competing background. Prior to the main experiments, each subject was presented with 150 repetitions of a short (100-msec with 10-msec on- and off-ramps) 1000-Hz tone pip to elicit the auditory N1 response. After confirming that N1 response was present for the subject, (s)he was tested with clean speech (quiet [Q] condition) and with speech mixtures (competing speech [CS] condition). The clean speech in Q conditions or the attended speech in the CS conditions was presented at 65 dB A. Each subject completed the Q conditions before the CS conditions. For the Q listening conditions (see Fig. 1 A), subjects were asked to either pay attention to the presented speech stimuli (active listening) while fixating their eyes on a crosshair on a computer screen in front of them, or not attend to the speech sounds but pay attention to a silent movie on a computer screen in front of them (passive listening). The silent movie was extracted from the animated film “Snow White.” Four randomly chosen 1-min speech segments (two from each speaker) were presented to each subject for each quiet listening condition. Each of the 1-min speech segments was presented 10 times, resulting in a total of four blocks of testing per listening condition. All subjects were tested with the active listening condition first, followed by the passive listening condition. The active and passive conditions were tested in two different 2-hour test sessions.

For the CS conditions (see Fig. 1B), each subject was presented with two randomly chosen 1-min speech mixtures for each TMR condition. The experiment was divided into four blocks per TMR, such that each of the two speech mixtures was used for two blocks. Each trial began with a written text cue (the word “male” or “female”) on a computer screen to indicate which speaker the subject should pay attention to in the mixture. The cue lasted for 1 sec and was the replaced by a crosshair, where subjects maintained visual fixation for the duration of the trial. Subjects were asked to focus on the male speaker in block 1 and the female speaker in block 2 for the first and second speech mixture, respectively. In block 3 and block 4, subjects switched their attention to the other speaker for the same speech mixtures as in block 1 and block 2, respectively. For each mixture, the unattended speaker started 1-sec after the attended speaker to help the subject listen to the correct speaker as cued. Similar to the Q conditions, each of the 1-min speech mixtures were presented 10 times per block. Blocks 1 and 2 for each of the two TMR conditions were tested in one 2-hour test session. Blocks 3 and 4 were tested in a different 2-hour session. The TMR of 6 dB (i.e., the level of the attended speech was 6 dB higher than that of the unattended speech) was tested first, followed by the 0 dB TMR condition for all subjects.

Subjects were asked to answer six true/false, multiple choice, or open-ended story comprehension questions for the attended speech after each test block for the active listening in the Q condition and for the CS conditions. The averaged comprehension accuracy for the attended speech was 93-94% across conditions (one-way repeated measures ANOVA, $P > 0.05$). These performance levels suggest that the subjects successfully attended to the target speech signal as instructed.

2.3. Data Recording and Processing

2.3.1. EEG recording and preprocessing—EEG signals were recorded using a 16-channel gUSBamp system with a.Butterfly electrodes from G.tec in a double-walled sound-isolated booth, at a sampling rate of 256 Hz. The placement of the electrodes followed the International 10/20 system (Oostenveld & Praamstra, 2001). A reference electrode was placed on the left earlobe. A bandpass filter between 0.1 and 100 Hz and a notch filter at 60 Hz were applied to the EEG recording online.

Due to the fact that the speech stimuli were not exactly equal in length (i.e., 1-min), subsequent analyses were performed on the first 50 sec of the ongoing neural responses to the speech signal (excluding the first sec after the onset of the speech to minimize the effect of the large onset response[1]). For each subject, the EEG signals were then filtered between 2 and 40 Hz and the stimulus phase-locked response component was extracted using denoising source separation (DSS; de Cheveigné & Simon, 2008). DSS is a blind source separation method, which extracts neural activity that is consistent over trials. The first DSS component was further analyzed, which had been previously shown to be effective in capturing envelope-tracking neural activity (Ding & Simon, 2012a). The topography of the first DSS component shows strong activation in Cz, C3, and C4, and is lateralized to the left hemisphere (Fig. 2), consistent with previous EEG studies on speech envelope tracking (Lalor et al., 2009; Power et al., 2012; Crosse & Lalor, 2014).

2.3.2. Speech envelopes—Temporal envelopes of the clean speech, as well as the individual attended and unattended speech that were used to make the speech mixtures were extracted using a Hilbert transformation. The speech envelopes were then resampled to a sampling rate of 256 Hz, followed by bandpass filtering between 2 to 40 Hz to match the sampling rate and the bandwidth of the EEG signals.

2.3.3. Cross-correlation analysis—From each subject and each test condition, the cross-correlation between the speech envelope and the EEG responses was calculated over the entire speech segment duration, i.e. 49 sec, for time lags varying from -200 to 600 msec at every 4-msec intervals. For the CS conditions, the cross-correlation functions were separately computed between the EEG signals and the temporal envelopes of the attended and unattended speech. Two factors can contribute to the polarity of the cross-correlation. One is the direction of the neural current and the other is whether the neural source is responsive to a power increase in the envelope or a power decrease. For example, a positive peak in the cross-correlation may indicate that a neural generator that produces a positive voltage on the scalp responds to a power increase in the speech envelope. Alternatively, it may also indicate a neural generator that produces a negative voltage on the scalp but tracks a power decrease in the envelope.

The group level cross-correlation function was computed by averaging the correlation efficient values across subjects for each time delay. Bootstrapping with 1,000 trials was

[1]The analysis procedure followed previous MEG studies (Ding & Simon, 2012a, 2012b). A separate analysis was applied, which further excluded the onset response to the unattended speech. That analysis yielded results essentially identical to the results reported here.

performed to establish the 95% confidence interval (CI) for cross-correlations between the EEG signal and its corresponding speech envelope. To estimate the chance-level cross correlation, we then separately calculated the 95% CI for random correlations for the active Q, passive Q, 6 dB TMR, and 0 dB TMR conditions, again using bootstrapping with 1,000 trials. For this purpose, correlations were computed between randomly selected segments of the EEG signals and the speech envelope at different time lags.

3. Results

Ongoing EEG responses were recorded from subjects listening to a narrated story presented either in quiet or in the presence of a competing talker of the opposite gender. Cross-correlation coefficient values were computed at different time lags. Figures 3, 4, and 5 show the cross-correlation functions for different listening conditions for attended and unattended speech. For each function, the solid lines represent the correlation coefficients at different time lags averaged across subjects. The shaded areas indicate 95% CI from bootstrapping. The dashed lines represent the 95% CI for random correlation. Tables 1 and 2 summarize the mean and the 95% CI for latency and peak correlation at time lags around 100 and 200-300 msec, respectively, for the attended and unattended speech at different listening conditions. The latencies and peak correlations were determined as the local minima and local maxima for the negative peak (N1) before 200 msec and the positive peak (P2) between 200 and 400 msec. Significant difference in latency or peak correlation is defined as non-overlapping 95% CIs between two listening conditions ($P < 0.05$).

3.1. Quiet conditions

When listening to speech in quiet, auditory cortical responses phase lock to the speech envelope, consistent with previous studies (Aiken & Picton, 2008, Ding & Simon, 2012a; Lalor & Foxe, 2010). Figure 3 shows the cross-correlation coefficients between EEG and the speech envelope at different time lags in the Q active and passive listening conditions. For active listening, the grand average cross-correlation function (across speaker gender, Fig. 3C) had a significant prominent positive peak correlation (r = 0.07) at a time lag of 152 msec, and two smaller negative peaks at time lags of 70 msec (r = -0.03) and 313 msec (r = -0.03). These findings are in agreement with those reported by Aiken & Picton (2008) for active listening in quiet. These peaks roughly correspond to the N1, P2, and N2 components from the auditory evoked potential literature (Horton et al., 2013), hence we refer to them as N1, P2, and N2 for the rest of this paper. The latency and peak correlation of N1 and P2 in the cross-correlation functions were not significantly different between the active and passive listening.

When the neural responses to the male and female speaker were analyzed separately, we found an insignificant ($P > 0.05$) trend for the N1 and P2 amplitude to be smaller in the passive listening for the male speaker (Fig. 3A). This trend was not observed for the female speaker (Fig. 3B). These results may suggest that that bottom-up attention does affect the modulation of neural responses. The female voice is more perceptually salient than the male voice, such that the female-spoken passage may have popped up more frequently during passive listening.

3.2. Competing speaker conditions

For the CS conditions, the shape of the cross-correlation function was remarkably different between the attended and unattended speech. The neural responses to the male and female speakers were averaged since no significant difference was seen between them ($P > 0.05$). Figure 4 shows cross-correlation coefficients between the EEG signals and the speech envelope at different time lags at 6 dB (right) and 0 dB TMR (left). For *attended* speech, the correlation functions showed a prominent N1 at time lag around 90 msec and a P2 around 180 msec. As seen in Table 1 and 2, the mean correlation coefficient and latency of these peaks were not significantly different (overlapping 95% CI, $P > 0.05$) between the two TMR conditions. For N1, the mean latency was 94 msec for both the 6 dB and 0 dB TMR condition, and the mean correlation coefficient was -0.04 and -0.06 for the 6 dB and 0 dB TMR, respectively. For P2, the mean latency was 176 msec at the 6 dB TMR and 184 msec at 0 dB TMR, and the mean correlation coefficient was 0.05 for both TMRs.

For *unattended* speech, the cross-correlation functions were similar between 6 dB and 0 dB TMRs, and the latencies and the strength of correlations were not significantly different between the two TMR conditions. However, these functions were different from those observed for attended speech at both earlier (< 100 msec) and later (> 100 msec) time lags. First, there was a large positive peak (P1) around 40-50 msec for both TMRs, which was not observed for the attended speech. Second, for both TMRs, the N1 was significantly delayed (120 – 130 msec) compared to the attended speech (94 msec). The N1 was significantly reduced (r = -0.03) compared to the attended speech (r = -0.06) at 0 dB TMR, but not at 6 dB TMR. That is, there was an insignificant ($P > 0.05$) trend that the difference in N1 amplitude between attended and unattended speech was smaller for 6 dB than for 0 dB TMR. Third, the P2 was significantly delayed (260 – 270 msec) and reduced (r = 0.02) compared to that of the attended speech (r = 0.05 at around 180 msec) for both TMRs.

To further understand the attentional effect and the effect of task demands on neural responses to attended and unattended speech, we replotted the correlation functions to compare the *attended* speech in the Q active listening condition and in the CS condition at 0 dB TMR (Fig. 5, left panel). The N1 amplitude was significantly greater and the N1 latency was significantly longer for the CS condition than for the Q condition. The P2 amplitude, however was similar between Q and CS conditions.

To reveal possible neural suppression of the unattended speaker, we also compared the neural responses to the *unattended* speech in the Q passive condition and in the CS condition at 0 dB TMR (Fig. 5, right panel). The P2 correlation was significantly reduced for the unattended speech in the CS condition than that for the passive condition. Neural responses to the unattended speech in the CS condition also showed an enhanced correlation at 50 msec compared to passive listening. Furthermore, the latencies of N1 and P2 were significantly delayed for the unattended speaker in the CS condition compared to the passive listening condition by 55 and 114 msec, respectively. These findings suggest a suppression mechanism for unattended sound in the presence of competitions within the auditory modality.

4. Discussion

In the present study, we investigated how top-down attention modulated cortical tracking of the speech envelope in different listening conditions. In the within-modality condition, competing information sources, i.e. two speech streams, were both presented auditorily. In the cross-modality condition, however, competing information sources, i.e. a speech stream and a silent movie, were presented separately via the auditory and visual modalities. In the within-modality condition, our results showed that attention differentiates the neural phase locking to the attended and the unattended speakers, by modulating not only the gain but also the shape of the cross-correlation between speech and the neural response. In contrast, in the cross-modality condition, neural tracking of speech was not significantly different with regard to the subjects' attentional state to the auditory input.

4.1 Attention Modulation of the N1 Response

We observed phase-locked neural responses to the envelope of continuous speech presented in quiet and to the attended speech stream presented in a competing background. The timing of the peak correlations corresponded to the classic N1-P2 evoked potential components, consistent with previous studies (Aiken & Picton, 2008; Lalor & Foxe, 2010; Ding & Simon, 2012a, 2012b; Horton et al., 2013). We found an enhanced N1 amplitude for the attended speech in the CS condition compared to the active Q condition, which is possibly due to the higher attentional load required in "cocktail-party like" listening situations. This result supports the idea that the selective attentional effect occurs at an early processing stage at around 100 msec after the onset of the auditory stimuli (Hillyard et al., 1973; Hansen & Hillyard, 1988; Alho et al., 1994; Melara et al., 2002; Choi et al., 2014). The early effect of attention was also evident in the differences in the cross-correlation function between the attended and unattended speech in the CS conditions.

We also observed a significant difference in N1 latency for the attended speech between the active Q and CS conditions, as well as for the unattended speech between the passive Q and CS conditions, consistent with previous MEG studies (Ding & Simon, 2012a). The prolonged N1 latency may reflect the greater challenge in the CS conditions compared to the Q listening conditions (Alain & Woods, 1994). The observation of the prolonged N1 latency for speech comprehension in a competing background is also in agreement with results observed for tone detection and perception of speech signals in noise (Billings et al., 2011).

4.2 Attentional Modulation of the P1 Component

A somewhat surprising finding in our study is that the P1 amplitude was significantly greater for the actively ignored competitor compared to other listening conditions (Q or attended speech in the CS conditions). This difference was also previously observed by Chait et al. (2010). Using a tone embedded in noise paradigm, Chait et al. (2010, Supplementary Fig. 1) showed a significantly larger response to the actively ignored tone at around 50 msec (M50) compared to the response to the tone in a passive task. There are a few possible explanations for the enhanced P1 for the ignored sound. One is that P1 represents an early processing stage that differentiates the two segregated streams. This may be related to the neurophysiological findings that attention enforces to align the high-

excitability phases of neuronal activity to the attended stimulus and the low-excitability phase to the unattended stimulus (Lakatos et al., 2013; Schroeder & Lakatos, 2009). For this interpretation, the N1 polarity and P1 polarity represent the high-excitability phase and the low-excitability phase, respectively.

A second possible explanation for the enhanced P1 response is that it reflects an interaction between the P1 and N1 components during attentional modulation. It is possible that the N1 response temporally overlaps with the P1 response and a reduction in the N1 amplitude makes the P1 amplitude appear to be stronger in the EEG recording. This idea is qualitatively illustrated in Figure 6, in which the P1 response is modeled to not be modulated by attention (e.g., Picton & Hillyard, 1974; Ding & Simon, 2012b) while the N1 and P2 responses are (e.g., Hillyard et al., 1973; Picton & Hillyard, 1974; Ding & Simon, 2012a, 2012b; Power et al., 2012; Horton et al., 2013; Choi et al., 2014; O'Sullivan et al., 2014). In this multi-component model, the cross-correlation between speech envelope and EEG responses is qualitatively modeled as the sum of 3 components, i.e. the P1, N1, and P2. The amplitude of the 3 components could be independently modulated by attention to simulate the cross-correlation function in different experimental conditions. Each component is modeled using a Gaussian function. The peak of the Gaussian function is at 50 msec, 100 msec, and 200 msec for the P1, N1, and P2 components, respectively. The standard deviation of the Gaussian function is 100 msec, 167 msec, and 167 msec for the P1, N1, and P2 components, respectively (Fig. 6A). The weight/amplitude of each component is then chosen to qualitatively simulate the cross-correlation for the active listening condition in quiet and the 0 dB TMR condition (Fig. 6A & 6B). The weight of P1 is not modulated by condition and is always 1.0. The weight of N1 is 1.3, 2.0, and 0.25 for the Q active listening condition, attended speech in the CS condition, and unattended speech in the CS condition, respectively, and the weight of P2 is 1.3, 1.3, and 0.125 for these three conditions. In other words, the N1 amplitude is enhanced for the attended speech and reduced for the unattended speech in the CS condition, compared with the active listening condition. The P2 amplitude is reduced for the unattended speech but not enhanced for the attended speech compared with the Q condition. The differences in the weighting of these three components result in the larger P1 observed in the cross-correlation function for unattended speech.

4.3 Suppression of the Actively Ignored Sounds

When multiple sound streams are presented in the auditory modality, the system can selectively process one sound source by either enhancing the responses to that sound or by suppressing the responses to other sounds or by using both strategies. The differences in both the *magnitude* and *timing* of the P1-N1-P2 complex observed here indicate separate active enhancement and suppression mechanisms for attended and unattended speech, respectively.

By comparing the cross-correlations for the speech stream during passive listening in quiet and the cross-correlations for the unattended speech in a competing background, we can isolate the neural suppression mechanism related to the presence of a competing speech stream. In other words, we probe how the neural processing of speech is affected by the listening background when the subjects are not involved in any task related to that speech

stream. The difference between the two listening conditions is that the listeners had to actively ignore the competing signal in CS conditions but not necessarily do so in the passive condition. As is shown in Fig. 5, the EEG response is indeed different for the speech stream in the passive listening in the quiet condition than for the unattended speech stream in the CS condition. The P2 amplitude was significantly reduced for the actively ignored speech stream in the CS condition compared to that for the unattended speech stream in the passive listening condition, while the P1 response is enhanced. Taken together, the differences in neural responses to the unattended speech between the passive listening in Q and in CS conditions support the existence of the suppression mechanism due to selective attention.

4.4 Attentional Selection Across Sensory Modalities

The present study did not find a significant difference in the cross-correlation function between active and passive listening in quiet. However, previous studies have shown that attention can modulate neural activity when the subjects pay attention to different sensory modalities. For example, in Picton & Hillyard (1974), listeners were asked to detect and count the number of clicks that were slightly lower in intensity compared to a standard in an active listening condition. In a passive listening condition, listeners were asked to read a book and to disregard as much as possible the ongoing auditory stimuli, similar to the passive listening condition in the present study. These authors found that there was a significant increase in N1 and P2 amplitude for the attended condition compared to the ignored condition. A possible explanation for this discrepancy is the level of task difficulty between our study and Picton & Hillyard (1974). Picton & Hillyard required listeners to detect a small difference in level of the deviant signals, a relatively more difficult task compared to active listening to robust continuous speech in quiet in the present study.

5. Summary

The present study investigated differential neural tracking of the attended and unattended speech streams in different listening conditions. Our work extended previous EEG studies (Power et al., 2012; Horton et al., 2013; O'Sullivan et al., 2014) by (1) investigating how the neural responses to the attended speech are differentiated from the neural responses to the unattended speech when competing sensory information is presented in the same or a different sensory modality; (2) investigating how the neural responses to attended/unattended speech stream are modulated when the listening background changes from being quiet to containing a competing speaker; and (3) investigating attentional modulation of speech streams presented diotically rather than dichotically.

First, it is demonstrated that top-down attention significantly modulates neural tracking of the speech envelope when two competing speakers are presented diotically. Importantly, the effects of attentional modulation strongly depend on the response latency (Figs. 4 & 5). Specifically, the cross-correlation with the attended speaker shows salient N1 and P2 responses, while the cross-correlation with the unattended speaker shows a salient P1 response. Second, neural tracking of a single speech stream presented in quiet is not significantly different whether the subjects pay attention to the speech (active listening) or to the silent movie (passive listening). Third, by comparing the responses to the attended

speech in the Q active listening condition with the responses to the attended speaker in the CS conditions, it is observed that the N1 response is enhanced and delayed in the CS conditions. By comparing the responses to the unattended speech in the Q passive listening condition with the responses to the unattended speaker in the CS condition, it is observed that the P2 response is attenuated while the P1 response is enhanced in the CS condition, suggesting the presence of active suppression mechanisms in more challenging listening conditions. Taken together, these results demonstrate that attentional modulation of speech tracking responses strongly depends on the task and top-down attention can both enhance the neural response to the attended sound stream and attenuate the neural responses to the unwanted sound stream.

Acknowledgments

This work was supported by NIH R01-DC-012300. We thank the two anonymous reviewers for their helpful comments.

References

Aiken SJ, Picton TW. Human cortical responses to the speech envelope. Ear & Hearing. 2008; 29:139–157. [PubMed: 18595182]

Alain C, Woods DL. Signal clustering modulates auditory cortical activity in humans. Perception & Psychophysics. 1994; 56:501–516. [PubMed: 7991348]

Alain C, Achim A, Richer F. Perceptual context and the auditory selective attention effect on event-related brain potentials. Psychophysiology. 1993; 30:572–580. [PubMed: 8248449]

Alho K, Teder W, Lavikainen J, Näätänen R. Strongly focused attention and auditory event-related potentials. Biological Psychology. 1994; 38:73–90. [PubMed: 7999931]

Bidet-Caulet A, Mikyska C, Knight RT. Load effects in auditory selective attention: evidence for distinct facilitation and inhibition mechanism. NeuroImage. 2010; 50:277–284. [PubMed: 20026231]

Billings CJ, Bennett KO, Molis MR, Leek MR. Cortical encoding of signals in noise: effects of stimulus type and recording paradigm. Ear & Hearing. 2011; 32:53–60. [PubMed: 20890206]

Chait M, de Cheveigné A, Poeppel D, Simon JZ. Neural dynamics of attending and ignoring in human auditory cortex. Neuropsychologia. 2010; 48:3262–3271. [PubMed: 20633569]

Choi I, Wang L, Bharadwaj H, Sinn-Cunningham BG. Individual differences in attentional modulation of cortical responses correlate with selective attention performance. Hearing Research. 2014; 314:10–19. [PubMed: 24821552]

Crosse MJ, Lalor EC. The cortical representation of speech envelope is earlier for audiovisual speech than audio speech. Journal of Neurophysiology. 2014; 111:1400–1408. [PubMed: 24401714]

de Cheveigne A, Simon JZ. Denoising based on spatial filtering. Journal of Neuroscience Methods. 2008; 171:331–339. [PubMed: 18471892]

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

Ding N, Simon JZ. Emergence of neural encoding auditory objects while listening to competing speakers. Proceedings of the National Academy of Sciences USA. 2012b; 109:11854–11859.

Ding N, Simon JZ. Cortical Entrainment to Continuous Speech: Functional Roles and Interpretations. Frontiers Human Neuroscience. 2014; 810.3389/fnhum.2014.00311

Hansen JC, Hillyard SA. Temporal dynamics of human auditory selective attention. Psychophysiology. 1988; 25:316–329. [PubMed: 3406331]

Hillyard SA, Hink RF, Schwent VL, Picton TW. Electrical signs of selective attention in the human brain. Science. 1973; 182:177–180. [PubMed: 4730062]

Horton C, D'Zmura M, Srinivasan R. Suppression of competing speech through entrainment of cortical oscillations. Journal of Neurophysiolology. 2013; 109:3082–3093.

Howard MF, Poeppel D. Discrimination of speech stimuli based on neuronal response phase patterns depends on acoustics but not comprehension. Journal of Neurophysiology. 2010; 104:2500–2511. [PubMed: 20484530]

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

Lakatos P, Karmos G, Mehta AD, Ulbert I, Schroeder CE. Entrainment of neuronal oscillations as a mechanism of attentional selection. Science. 2008; 320:110–113. [PubMed: 18388295]

Lakatos P, Musacchia G, O'Connel MN, Falchier AY, Javitt DC, Schroeder CE. The spectrotemporal filter mechanism of auditory selective attention. Neuron. 2013; 77:750–761. [PubMed: 23439126]

Lalor EC, Foxe JJ. Neural responses to uninterrupted natural speech can be extracted with precise temporal resolution. European Journal of Neuroscience. 2010; 31:189–193. [PubMed: 20092565]

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

Luo H, Poeppel D. Phase patterns of neuronal responses reliably discriminate speech in human auditory cortex. Neuron. 2007; 54:1001–1010. [PubMed: 17582338]

Melara RD, Rao A, Tone Y. The duality of selection: excitatory and inhibitory processes in auditory selective attention. Journal of Experimental Psychology: Human Perception & Performance. 2002; 28:279–306. [PubMed: 11999855]

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

Näätänen, R. Attention and brain function. Hillsdale, NJ, USA: Lawrence Erlbaum Associates; 1992.

Neelon MF, Williams J, Garell PC. The effects of auditory attention measured from human electrocorticograms. Clinical Neurophysiology. 2006; 117:504–521. [PubMed: 16458596]

Oldfield RC. The assessment and analysis of handedness: the Edinburgh inventory. Neuropsychologia. 1971; 9:97–113. [PubMed: 5146491]

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

O'Sullivan JA, Power AJ, Mesgarani N, Rajaram S, Foxe JJ, Shinn-Cunningham BG, Slaney M, Shamma SA, Lalor EC. Atttentional selection in a cocktail party environment can be decoded from single-trial EEG. Cerebral Cortex. 2014 [Epub ahead of print]. 10.1093/cercor/bht355

Pasely BN, David SV, Mesgarani N, Flinker A, Shamma SA, Crone NE, Knight RT, Chang EF. Reconstructing speech from human auditory cortex. PLoS Biology. 2012; 10:e1001251. [PubMed: 22303281]

Picton TW, Hillyard SA. Human auditory evoked potentials. II: Effects of attention. Electroencephalography and Clinical Neurophysiology. 1974; 36:191–199. [PubMed: 4129631]

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. European Journal of Neuroscience. 2012; 35:1497–1503. [PubMed: 22462504]

Rif J, Hari R, Hämäläinen MS, Sams M. Auditory attention affects two different areas in the human supratemporal cortex. Electroencephalography and Clinical Neurophysiology. 1991; 79:464–472. [PubMed: 1721574]

Ross B, Hillyard SA, Picton TW. Temporal Dynamics of Selective Attention during Dichotic Listening. Cerebral Cortex. 2010; 20:1360–1371. [PubMed: 19789185]

Schroeder CE, Lakatos P. Low-frequency neural oscillations as instruments of sensory selection. Trends in Neuroscience. 2009; 32:9–18.

Snyder JS, Alain C, Picton TW. Effects of attention on neuroelectric correlates of auditory stream segregation. Journal of Cognitive Neuroscience. 2006; 18:1–13. [PubMed: 16417678]

Teder-Sälejärvi WA, Hillyard SA, Röder B, Neville HJ. Spatial attention to central and peripheral auditory stimuli as indexed by event-related potentials. Cognitive Brain Research. 1999; 8:213–227. [PubMed: 10556600]

Zion Golumbic E, Cogan GB, Schroeder CE, Poeppel D. Visual input enhances selective speech envelope tracking in auditory cortex at a “cocktail party”. Journal of Neuroscience. 2013; 33:1417–1426. [PubMed: 23345218]

Abbreviations

ANOVA	analysis of covariance
CI	confidence interval
CS	competing speaker
DSS	denoising source separation
EEG	electroencephalography
Hz	hertz
kHz	kilohertz
MEG	magnetoencephalography
msec	millisecond
Q	quiet
RMS	root-mean-square
TMR	target-to-masker ratio

Highlights

- Attention enhances responses to attended speech and suppresses ignored speech.
- Attentional modulation occurs at both early and later processing stages.
- Attentional effects differ depending on selection within or between modalities.
- Modulation of N1 responses is affected by attentional load and task demand.

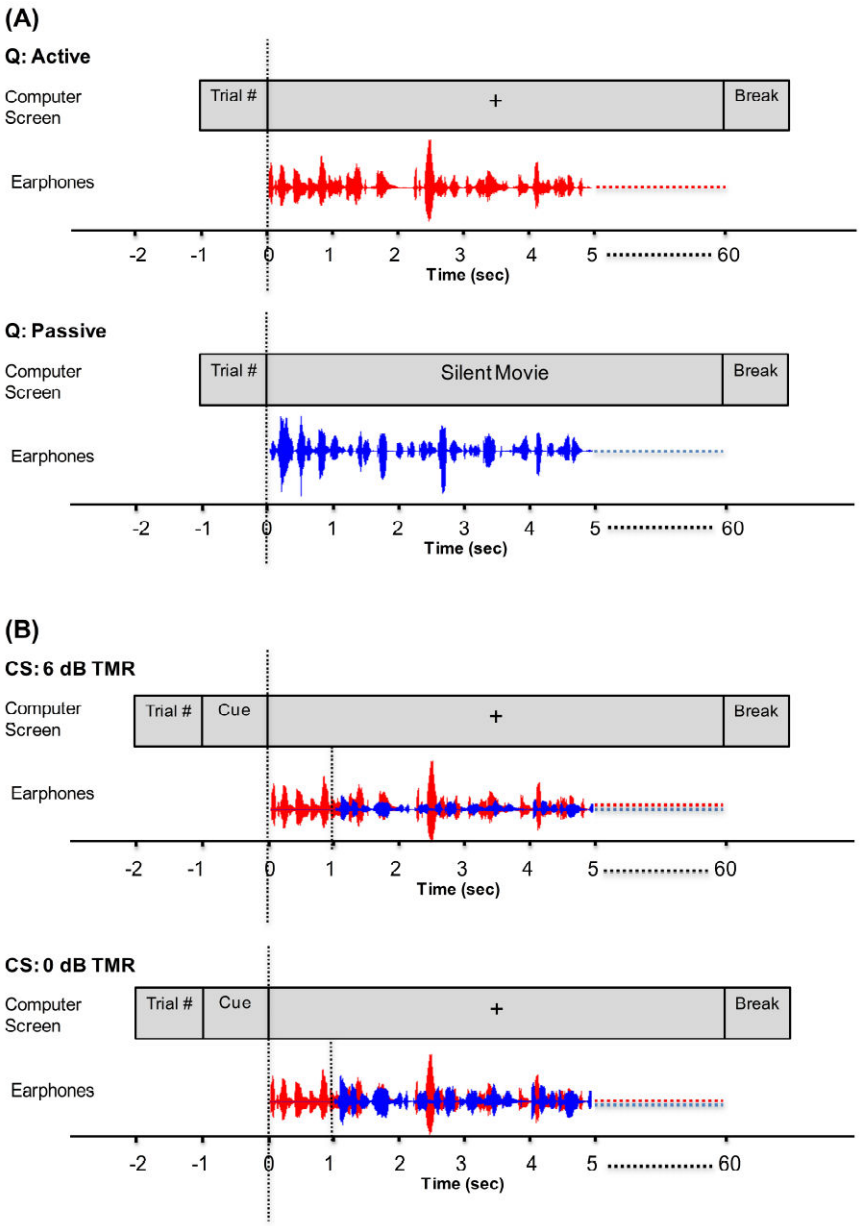

Figure 1.

Schematic illustration of the timing of the visual and auditory displays in one trial for each of the four testing conditions. Trial number indicates the number of repetitions of the sentence within each block. For Quiet conditions (panel A), the visual signal was either a fixation point "+" or a silent movie lasted for the duration of the auditory signal for the Active and Passive listening condition, respectively. For the CS conditions (panel B), a visual word cue ("male" or "female") was displayed for one second followed by a fixation point "+" lasted for the duration of the auditory signal. The auditory signal started after the end of the visual word cue. The unattended speech (blue) started one second after the attended speech (red). The unattended speech was the same intensity as the attended speech in the 0 dB TMR condition, but was 6 dB weaker in the 6 dB TMR condition.

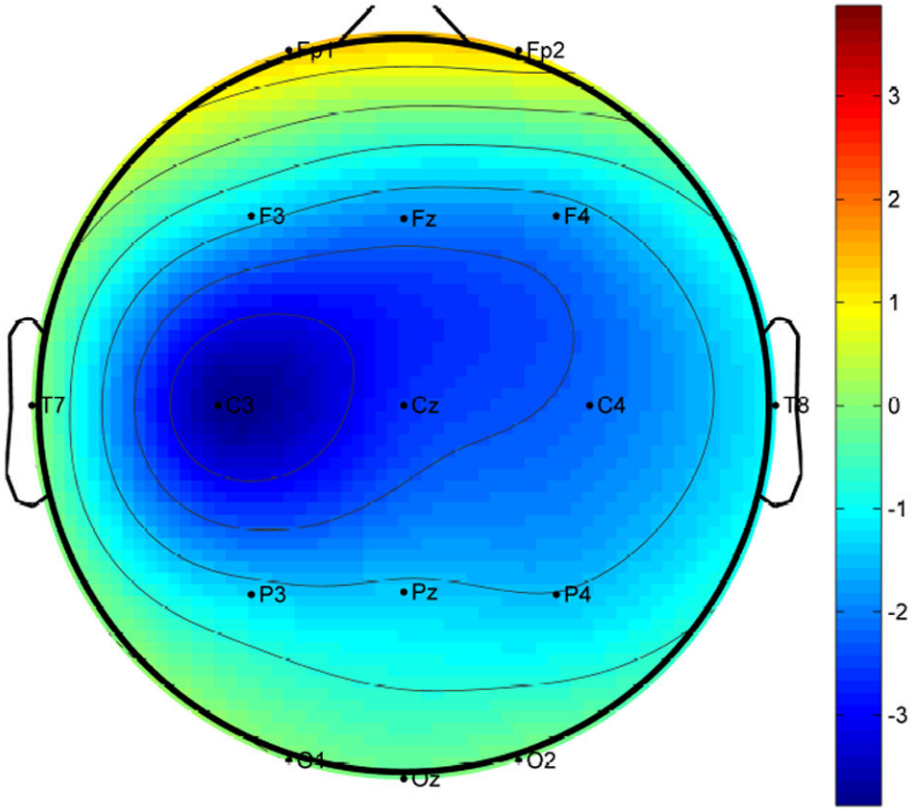

Figure 2.
Topography of the spatial filter coefficients for the first DSS component averaged across subjects.

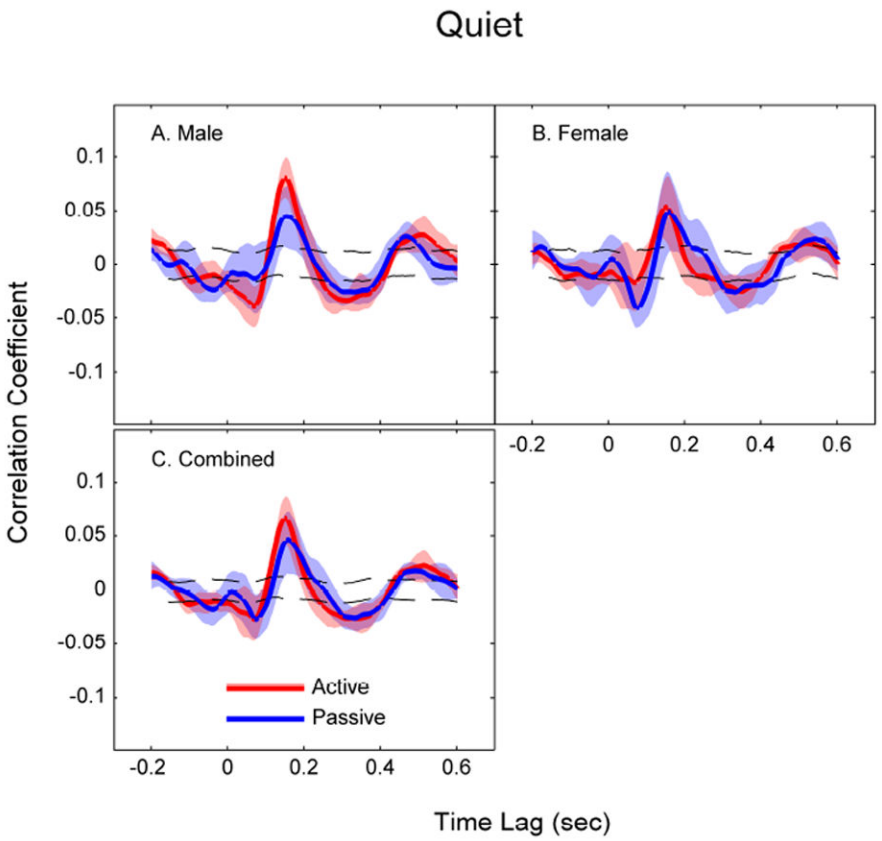

Figure 3.

Correlation coefficients between the EEG and the speech envelope at different time lags in the quiet conditions (red: active listening; blue: passive listening). Shaded areas denote 95% confidence interval (CI) of the cross-correlations. Dashed lines represent the 95% CI for random correlation. Cross-correlations were calculated separately for the male speaker (panel A), female speaker (panel B), as well as for both speakers combined (panel C).

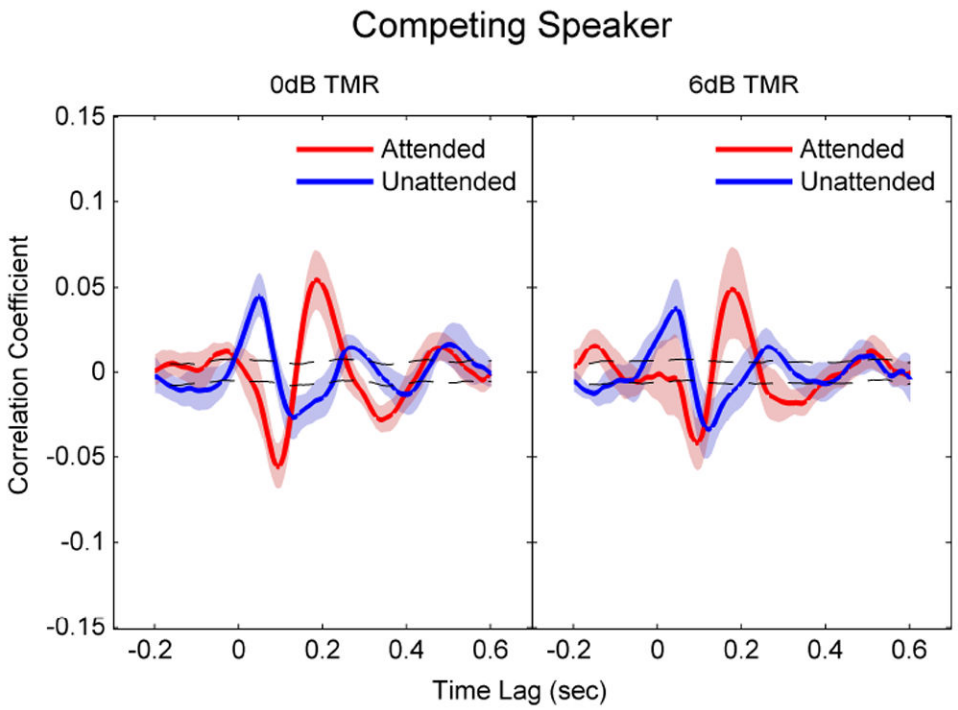

Figure 4.

Correlation coefficients between the EEG and the speech envelope for the attended and unattended speech at different time lags in the CS conditions (left: 0 dB TMR; right: 6 dB TMR). Shaded areas denote 95% confidence interval (CI) of the cross-correlations. Dashed lines represent the 95% CI for random correlation.

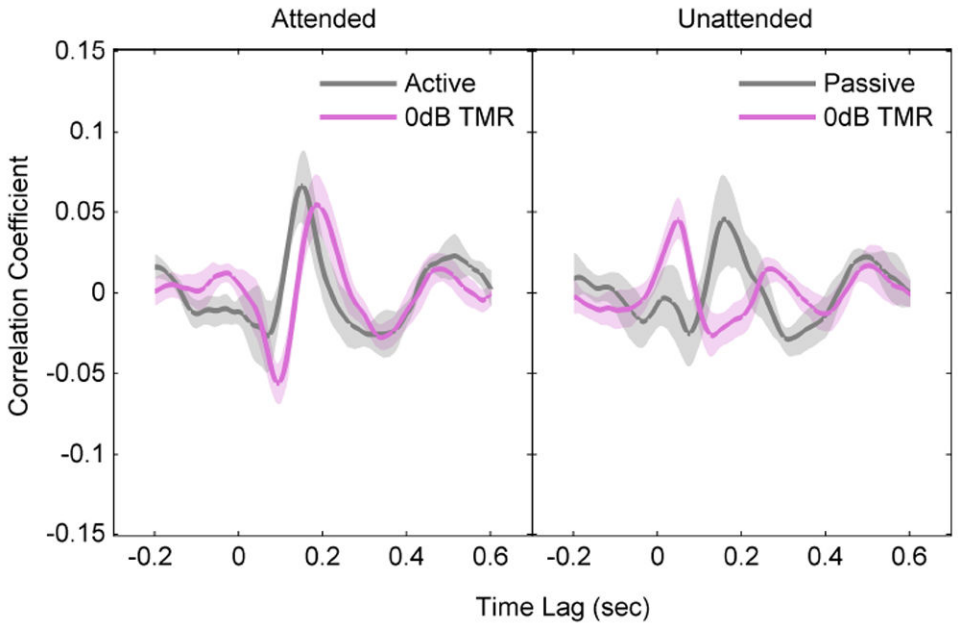

Figure 5.

Correlation coefficients between the EEG and the speech envelope at different time lags. Left panel compares cross-correlations for the attended speech for the Q and CS conditions. Right panel compares the results for the unattended speech. Shaded areas denote 95% confidence interval (CI) of the cross-correlations.

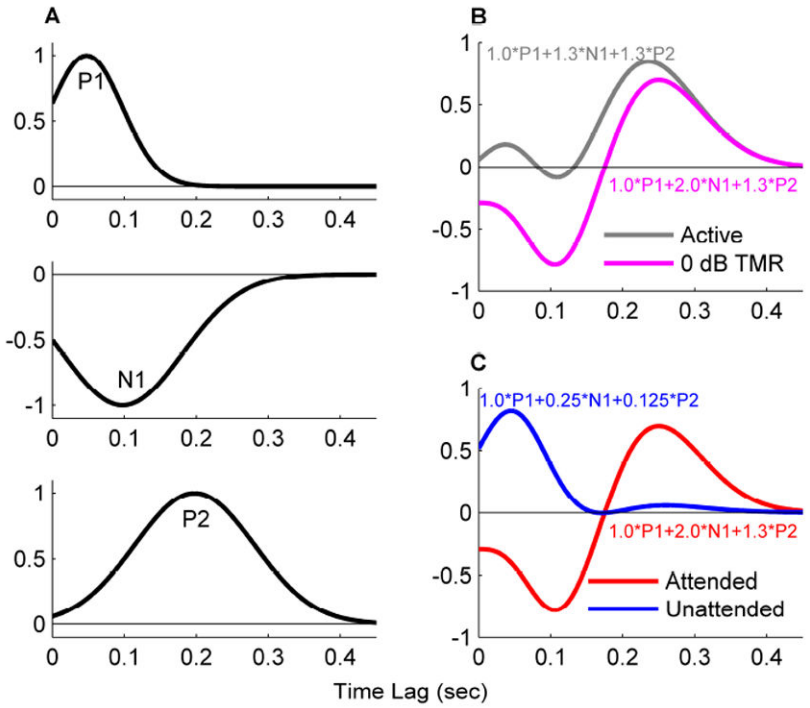

Figure 6.

A multi-component model for the cross-correlation between speech envelope and EEG. (A) The model consists of 3 components, i.e. P1, N1, and P2. The cross-correlation function is modeled as the weighted sum of the 3 components. (B) Modeling the response during active listening and for the attended speech in the 0 dB TMR condition (cf. Figure 5 left). The amplitude of P1 and P2 remains the same in these two conditions, while the amplitude of N1 is bigger for 0 dB TMR condition. (C) Modeling the responses to the attended and unattended speech in the 0 dB TMR condition (cf. Figure 4 left). The amplitude of P1 remains the same in both conditions. The amplitude of N1 and P2, however, is bigger for attended speech. The weights of each component of the modeled responses are indicated in panels B and C.

Table 1

The latency of the peaks in the cross-correlation functions for the attended and unattended speech in different listening conditions.

Attention	Peak	Listening condition	Mean (msec)	95% CI (msec)
Attended speech	N1	Active	70	0 — 78
		CS at 6dB TMR	94	82 — 106
		CS at 0dB TMR	94	86 — 102
	P2	Active	152	145 — 156
		CS at 6dB TMR	176	164 — 195
		CS at 0dB TMR	184	176 — 203
Unattended speech	N1	Passive	74	0 — 101
		CS at 6 dB SNR	121	102 — 191
		CS at 0 dB SNR	129	121 — 195
	P2	Passive	156	148 — 221
		CS at 6 dB SNR	262	250 — 305
		CS at 0 dB SNR	270	254 — 309

Table 2

The maximum correlation value of the peaks in the cross-correlation functions for the attended and unattended speech in different listening conditions.

Attention	Peak	Listening condition	Mean (*r*)	95% CI (*r*)
Attended speech	N1	Active	-0.03	-0.01 — -0.05
		CS at 6dB TMR	-0.04	-0.03 — -0.06
		CS at 0dB TMR	-0.06	-0.05 — -0.07
	P2	Active	0.07	0.04 — 0.09
		CS at 6dB TMR	0.05	0.03 — 0.07
		CS at 0dB TMR	0.05	0.04 — 0.07
Unattended speech	N1	Passive	-0.03	-0.01 — -0.05
		CS at 6 dB SNR	-0.03	-0.02 — -0.05
		CS at 0 dB SNR	-0.03	-0.02 — -0.04
	P2	Passive	0.05	0.02 — 0.07
		CS at 6 dB SNR	0.02	0.01 — 0.03
		CS at 0 dB SNR	0.02	0.01 — 0.02