International Journal of Bioelectromagnetism
Vol. 13, No. 4, pp. 217-222, 2011

EEG reveals divergent paths for speech envelopes during selective attention

Cort Horton[a], Michael D'Zmura[a], and Ramesh Srinivasan[a,b]
[a] *Dept. of Cognitive Sciences, University of California, Irvine, USA*
[b] *Dept. of Biomedical Engineering, University of California, Irvine, USA*
Correspondence: Cort Horton, Dept. of Cognitive Sciences,
University of California, Irvine, USA, CA 92697-5100
E-mail: chorton@uci.edu

Abstract. This paper reports the results of an experiment that was designed to determine the effects of attention on the representation of speech signal envelopes in EEG. We recorded EEG while subjects attended to one of two concurrently presented and spatially separated speech streams. Cross-correlating the speech envelope with signals from individual EEG channels reveals clear differences between the neural representations of the envelopes of attended and unattended speech. The two concurrently presented speech signals were amplitude modulated at 40 and 41 Hz, respectively, in order to investigate the effects of attention on speech signal gain. The modulations elicited strong steady-stage responses that showed no effects of attention. We conclude that the differences between the representations in EEG of the envelopes of attended and unattended speech reflect some form of top-down attentional control.

Keywords: EEG; Speech Perception; Selective Attention; Steady-State Responses

1. Introduction

Humans are skilled at tracking the voice of one among many speakers, often referred to as the "cocktail party effect" [for review see Bronkhorst, 2000]. The pattern of amplitude modulations found in speech, known as the envelope, may be of particular importance for this ability. The envelope represents the slow (mostly under 20Hz) changes in amplitude over time scales that correspond to phonemes and syllables [Rosen, 1992].

Researchers studying audition have known for quite some time that neural populations can track amplitude modulations in the frequency range of envelopes [Picton et al., 1987]. In addition, EEG and MEG studies of speech perception have repeatedly noted the importance of low frequency activity, especially in the theta band (4-8 Hz), for the comprehension of sentences and the discrimination between their evoked neural signals [Ahissar et al., 2001; Suppes et al., 1998; Luo and Poeppel, 2007]. Those frequencies correspond to the frequencies in speech envelopes that are most important for intelligibility [Drullman et al., 1994]. Motivated by these studies, Aiken and Picton [2008] looked for EEG signals that track complex speech envelopes and found dipole sources in auditory cortex that are significantly correlated to envelopes at a delay of 180ms. While this provided the first evidence that neural populations maintain a representation of speech envelopes, the authors could not ascertain from their data whether the representation is the result of a top-down mechanism that actively tracks the envelope, or simply a bottom-up signal that reacts to changes in the envelope. Investigating that question is the primary purpose of the present study.

In a study of the auditory system's response to a variety of stimuli, Deng and Srinivasan [2010] found signals in EEG that were correlated to the envelope of speech, but not to that of reversed speech. This suggests that envelope representations need not be purely stimulus-driven, as reversed speech shares the same spectral and temporal content as normal speech. In what may be the most directly applicable study, Kerlin et al. [2010] used a selective attention paradigm to search for EEG activity that could discriminate which of two sentences was being attended. Their findings suggest the importance of information from lower frequency bands, and they argued that selective attention depends on a frequency-specific increase in the response of auditory cortical areas to the attended speaker.

Our goal in this study was to investigate the source and functional significance of the neural representation of speech envelopes. We recorded EEG during a selective attention experiment that incorporates several improvements over previous methods. The results demonstrate that neural representations of the envelopes of attended and unattended speech differ in ways that cannot be explained as a simple difference in gain.

2. Methods

2.1. Subjects and Design

Nine subjects (1 female) aged 21-29 participated in the study. Subjects were seated in a sound-attenuating chamber with a computer monitor 1.5 meters ahead. A pair of FinalSound 400i electrostatic loudspeakers was positioned 45° to each side of the monitor, also at a distance of 1.5 meters from the subject. These speakers were selected due to their excellent temporal and frequency response and lack of electromagnetic interference with EEG recording. At the start of each trial, the monitor displayed a prompt to attend to either the left or right speaker, based on a fully randomized design. The prompt was followed by the appearance of a small cross, on which subjects were instructed to maintain fixation for the duration of the trial. One second after the cross appeared, the loudspeakers began playing speech stimuli constructed for the trial (details below), which lasted for 22 to 25 seconds. One second after the stimuli finished, the monitor displayed the transcript of a randomly chosen sentence from the trial. Subjects had to indicate with a button press whether the sentence had been played through the prompted speaker. This task was designed so that successful attention to the prompted speaker would result in a mildly challenging recognition task, while attempting to listen to both speakers and then deciding post hoc on the source of the test sentence would be very difficult. Subjects completed 320 trials, broken into 8 separate 20 minute sessions that were typically collected over the space of two weeks.

2.2. Stimuli

Trial stimuli were concatenations of sentences drawn randomly from the TIMIT speech corpus [Garofolo et al., 1993], which contains thousands of sentences recorded by both male and female speakers from a variety of dialect regions across the United States. Two such stimuli were played simultaneously through independent left and right audio channels. For each channel's stimulus, we continued to draw sentences from the corpus until the stimulus was at least 22 seconds long, which typically required the concatenation of seven or eight sentences. The shorter stimulus was then appended with zeros to match the duration of the longer; these together provided a single stereo stimulus that ranged from 22 to 25 seconds long. No sentence was ever reused within an experimental session. In order to minimize intensity differences among sentences, each sentence was treated with a subtle dynamic range compression and normalized to have equal RMS power. The volume of the loudspeakers was set so that the mean sound pressure level measured at the position of the subject was 65 dB_{SPL}.

The left and right channel stimuli were sinusoidally amplitude modulated at 40 and 41 Hz, respectively, in order to induce steady-state responses in neural populations. These frequencies were chosen because they induce relatively large steady-state responses to auditory stimuli [Picton et al., 2003] and are well above the peak modulation frequencies of speech envelopes [Rosen, 1992]. The amplitude modulation is perceived as a roughness similar to someone speaking through a fan, and only slightly reduces the intelligibility of the speech.

2.3. EEG Recording and Pre-Processing

EEG data were recorded from 128 channels using an amplifier, electrode caps, and digitization software made by Advanced Neuro Technology®. The EEG was sampled at 1024 Hz with an online average reference and imported into Mathworks MATLAB® for all further offline analyses. The EEG was filtered with a pass band of 1 to 50 Hz using zero-phase Butterworth filters and then down-sampled to 256 Hz. One subject's EEG was visibly contaminated on most trials by a variety of movement-related artifacts; these data were excluded from all further analysis. All data analyses were constrained to a window starting 1 second after the onset of the stimuli and ending 20 seconds later. This ensured that the large onset responses which sometimes dominated other studies' results [e.g. Aiken and Picton, 2008; Kerlin et al., 2010] were not included.

Although subjects were instructed to minimize movements during trials, with long trials some blinks were inevitably recorded in every subject's EEG. Independent Component Analysis was used on each subject's EEG to identify and remove activity related to blinks [Jung et al., 2000]. This resulted in the removal of 1-2 components for each subject that showed the distinctive frontal topography and spectral content indicative of blinks. This correction had no effect on any of the following results.

2.4. Analysis of Envelope-EEG Cross-Correlation

We used a cross-correlation method similar to that of Aiken and Picton [2008] to find activity in EEG channels that represented the stimuli envelopes. This method measures the similarity of two signals as a function of lag. If an EEG channel records signals from neural population which follow an

envelope at some latency, then one should find a deviation from zero in the cross-correlation function at a lag equal to that latency. Using this method, we can extract an evoked response to the speech envelope. On each trial, we cross-correlated three different envelopes with each channel of EEG: the attended and unattended envelopes (which are uncorrelated), and a control envelope. This control is the envelope of a stimulus selected randomly from a different trial and is used to estimate the amount of cross-correlation that occurs due to chance. The outputs of the cross-correlations are binned by condition, averaged over trials, and then averaged across subjects.

2.5. Analysis of Steady-State Auditory Evoked Responses

Trial data were transformed into the frequency domain using MATLAB's Fast Fourier Transform. The Fourier coefficients corresponding to the amplitude modulation frequencies 40 and 41Hz were first binned by condition and averaged over trials for each subject. Fourier coefficients contain information about both the amplitude and phase of responses. Yet the phase of steady-state auditory responses can vary greatly among subjects and is not relevant to our analysis [Picton et al., 2003]. To compensate for this, for each subject we calculated the mean phase of the steady-state responses at the frontal-central cluster of electrodes where the responses are maximal. We then rotated the phase of every channel by the mean phase so that the cluster now had a mean phase of zero while preserving phase differences among channels. After rotating each subject's coefficients in this way, all subjects' steady-state responses showed very similar phase patterns across the head; these were then averaged across subjects.

3. Results

3.1. Behavioral Data

Subjects performed well on the task, with a mean accuracy of 83.3% (SD 5.1%), with a slight increase in performance on right trials as compared to left (81.9% left, 84.8% right). At the end of the experiment, most subjects reported that their errors were primarily due to memory constraints, rather than lapses in attention. Although anecdotal, this account is in agreement with the observation that all analyses show no differences between results calculated using all trials vs. correct trials only.

3.2. Envelope-EEG Cross-Correlation

The average cross-correlations for each channel of EEG are plotted separately for attended and unattended envelopes in Figure 1. Both plots have peaks in correlation strength at multiple delays, indicating that a representation of the envelopes may persist across multiple steps of processing in the speech pathway. The control condition (not plotted) had none of the structure seen in the attended and unattended cases. The dashed lines on the attended and unattended plots indicate the maximum correlation recorded in the control condition, which is a good estimate of the largest correlation we would expect to see by chance. Based on that criterion, both conditions show correlations far above chance, with the attended case exhibiting stronger correlations in general. Note that the magnitudes of the correlations reported here are relatively small because the stimulus-related activity is only a small part of the EEG being recorded at each channel. This should not be confused with a weak effect size.

The attended envelope cross-correlation shows four major peaks, of which only the first has a clear counterpart in the unattended envelope. That first peak, occurring at a very short delay, shows the vertex maximum typical of very early auditory evoked potentials. This may reflect sub-cortical and very early cortical responses to the envelope modulations. The later peaks in the attended cross-correlation function show more complex topographies, with patterns of positive and negative correlations that are consistent with the types of dipole sources that EEG and MEG studies frequently localize to superior temporal cortex [e.g. Herdman et al., 2002; Aiken and Picton, 2008]. Overall, the pattern of correlations with the attended envelope agrees with and expands upon the results from Aiken and Picton [2008].

The results suggest that unattended envelopes have a less robust neural representation than attended envelopes after the early peak, although they still show a fairly large correlation at a 220 msec delay that does not match the topography of corresponding attended peak. Taken as a whole, these cross-correlations suggest that the neural envelope representations of attended and unattended speech differ in more fundamental ways than simply gain.

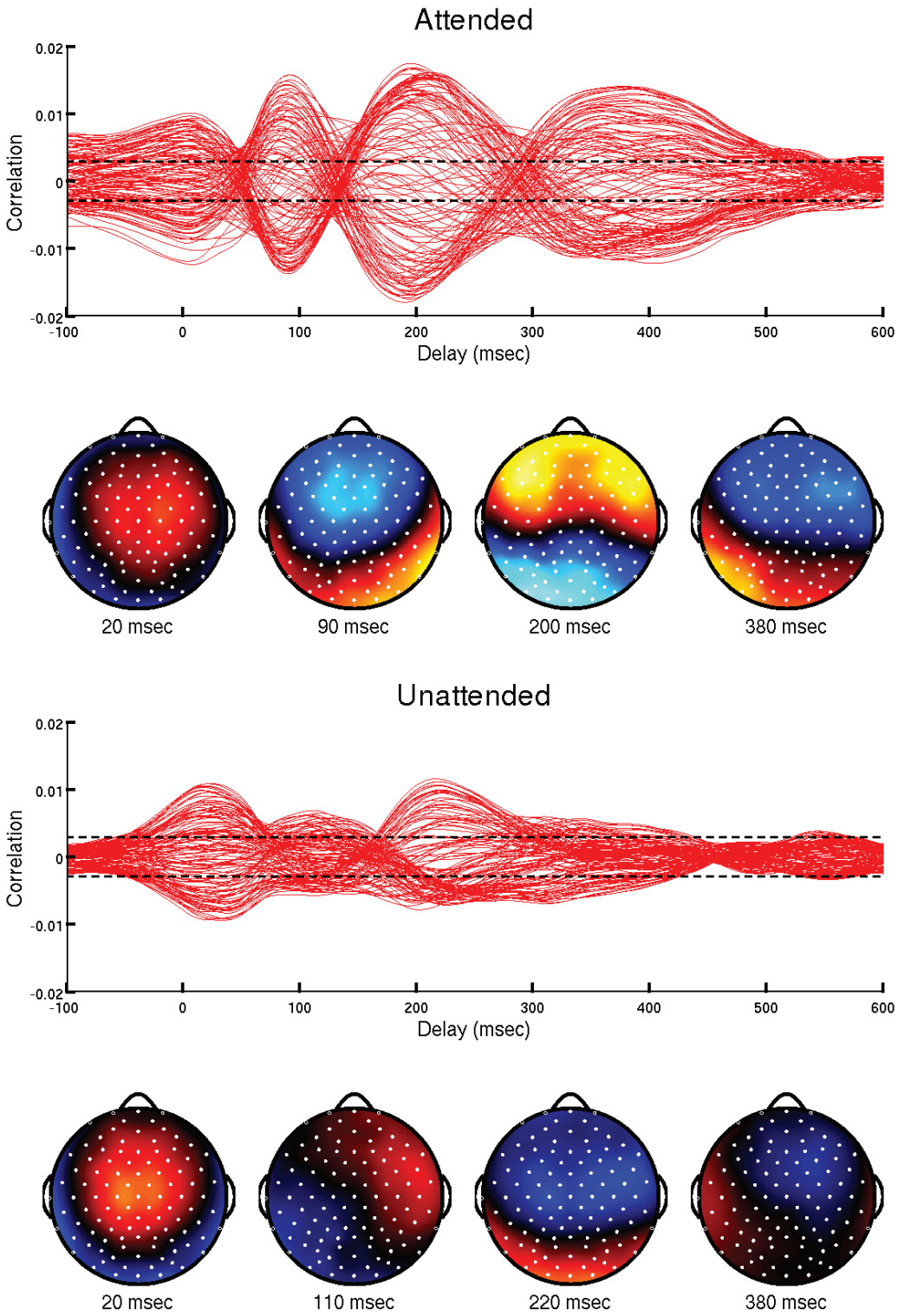

Figure 1. *Envelope-EEG cross-Correlations. Plots of the average correlation between each channel of EEG with the stimulus envelopes as a function of lag. Each trace represents a separate channel of EEG. The dashed line is set to the maximum value observed in the control condition, and represents the largest correlations we expect to occur by chance. Peaks in the cross-correlation functions are further illustrated with topographic plots underneath. Warm colors denote positive correlations, while cool colors denote negative correlations.*

3.3. Steady-State Auditory Evoked Responses

The amplitude modulation elicited robust steady-state responses at 40 and 41 Hz. Figure 2 shows the typical topography of these responses, as well as the average power spectrum for a channel over the frontal maximum. The scalp distribution seen here, including the reversal of phase between the frontal and occipital maxima, is typical of auditory steady-state studies, and it thought to reflect a mixture of sources from midbrain, thalamic, and cortical tissues [Herdman et al., 2002]. We found no effect of attention. All four steady-state responses (left attended, left unattended, right attended, right unattended) were indistinguishable from one another in both amplitude and phase.

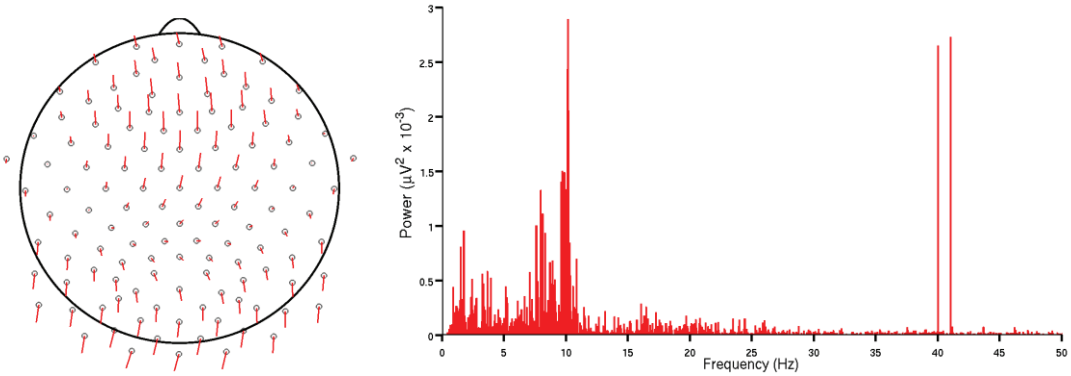

Figure 2. *Steady-state responses. Left: topographic plot displaying the average steady-state responses recorded at each EEG channel. The length of the line represents amplitude and the direction indicates phase. Right: average power spectrum at channel Fz for a representative subject. Steady state responses are visible as large peaks in power at 40 and 41 Hz.*

4. Discussion

We have shown evidence that the neural responses to attended and unattended speech differ after the initial onset response. The differences are not attributable to a change in gain of the sort suggested in Kerlin et al. [2010] because the timing and topography of peaks in the cross-correlation functions do not match. Furthermore, the lack of an effect of attention on steady-state responses speaks against any generalized gain-based account of the differences.

Two accounts are consistent with the data. In the first, the components seen in the cross-correlation functions that are unique to attended speech represent the top-down signals that have been postulated in previous studies. These signals could be involved in a variety of functions, such as chunking speech streams into syllable-sized pieces, or selecting speech for further processing in a noisy situation. A logical next step under this account would be to investigate the function and attempt to modulate the separate peaks identified by this cross-correlation procedure. The second account is that the signals reflected in the attended envelope cross-correlation are, in fact, stimulus-driven, and that the differences between conditions are due to some top-down signal that suppresses activity from unattended sources beyond the initial onset response. The end result of both accounts is the same--attended signals get processed, unattended ones do not--but the mechanisms are very different. Future studies are needed to tease these accounts apart.

Finally, these methods could provide useful insights into the debate in the speech perception literature regarding the hemispherically asymmetric processing of speech [e.g. Obleser et al., 2008]. Although not examined here in detail, the 90 and 380 ms responses to attended speech seem to be strongly lateralized. Future studies could systematically modulate certain spectral and temporal aspects of speech stimuli to investigate the effects of those modulations on these components.

References

Ahissar E, Nagarajan S, Ahissar M, Protopapas A, Mahncke H, Merzenich MM. Speech comprehension is correlated with temporal response patterns recorded from auditory cortex. *Proc. of the National Academy of Sciences*, 98: 13367-13372, 2001.

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

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

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

Drullman R, Festen JM, Plomp R. Effect of temporal envelope smearing on speech reception. *Journal of the Acoustical Society of America*, 95: 1053-1064, 1994.

Garofolo J, Lamel L, Fisher W, Fiscus J, Pallett D, Dahlgren N, Zue V. TIMIT acoustic-phonetic continuous speech corpus. 1993.

Herdman AT, Lins O, Van Roon P, Stapells DR, Scherg M, Picton TW. Intracerebral sources of human auditory steady-state responses. *Brain Topography*, 15(2): 69-86, 2002.

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

Kerlin JR, Shahin AJ, Miller LM. Attentional gain control of ongoing cortical speech representations in a "cocktail party". *The Journal of Neuroscience*, 30(2): 620-628, 2010.

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

Obleser J, Eisner F, Kotz S. Bilateral speech comprehension reflects differential sensitivity to spectral and temporal features. *The Journal of Neuroscience*, 28(32): 8116-8124, 2008.

Picton TW, Skinner CR, Champagne SC, Kellet AJC, Maiste AC. Potentials evoked by the sinusoidal modulation of the amplitude or frequency of a tone. *The Journal of the Acoustic Society of America*. 82(1): 165-178, 1987.

Picton TW, John MS, Dimitrijevic A, Purcell D. Human auditory steady-state responses. *International Journal of Audiology*. 42: 177-219, 2003.

Rosen S. Temporal information in speech: Acoustic, auditory and linguistic aspects. *Philosophical Transactions: Biological Sciences,* 336(1278): 367-373, 1992.

Suppes P, Han B, Lu ZL. Brain-wave recognition of sentences. *Proc. of the National Academy of Sciences*, 95: 15861-15866, 1998.