Envelope responses in single-trial EEG indicate attended speaker in a ‘cocktail party’

2014 J. Neural Eng. 11 046015

(http://iopscience.iop.org/1741-2552/11/4/046015)

IOP Publishing Journal of Neural Engineering
J. Neural Eng. 11 (2014) 046015 (12pp) doi:10.1088/1741-2560/11/4/046015

Envelope responses in single-trial EEG indicate attended speaker in a 'cocktail party'

Cort Horton[1], Ramesh Srinivasan[1,2] and Michael D'Zmura[1]

[1] Department of Cognitive Sciences, University of California, Irvine, CA 92697, USA
[2] Department of Biomedical Engineering, University of California, Irvine, CA 92697, USA

E-mail: chorton@uci.edu

Received 20 May 2014
Accepted for publication 30 May 2014
Published 25 June 2014

Abstract

Objective. Recent studies have shown that auditory cortex better encodes the envelope of attended speech than that of unattended speech during multi-speaker ('cocktail party') situations. We investigated whether these differences were sufficiently robust within single-trial electroencephalographic (EEG) data to accurately determine where subjects attended. Additionally, we compared this measure to other established EEG markers of attention. *Approach.* High-resolution EEG was recorded while subjects engaged in a two-speaker 'cocktail party' task. Cortical responses to speech envelopes were extracted by cross-correlating the envelopes with each EEG channel. We also measured steady-state responses (elicited via high-frequency amplitude modulation of the speech) and alpha-band power, both of which have been sensitive to attention in previous studies. Using linear classifiers, we then examined how well each of these features could be used to predict the subjects' side of attention at various epoch lengths. *Main results.* We found that the attended speaker could be determined reliably from the envelope responses calculated from short periods of EEG, with accuracy improving as a function of sample length. Furthermore, envelope responses were far better indicators of attention than changes in either alpha power or steady-state responses. *Significance.* These results suggest that envelope-related signals recorded in EEG data can be used to form robust auditory BCI's that do not require artificial manipulation (e.g., amplitude modulation) of stimuli to function.

Keywords: selective attention, speech envelopes, brain–computer interfaces, alpha lateralization, steady-state responses

(Some figures may appear in colour only in the online journal)

1. Introduction

A great deal of effort has been devoted to mapping out the relationship between the acoustic properties of speech utterances and their associated neural responses. The feature of speech with the strongest representation in the cortex appears to be its temporal envelope. Researchers have found strong correlations between auditory cortical activity and speech envelopes [1–3] which appear to be produced by the synchronization (or 'phase-locking') of endogenous oscillations to the slow (<10 Hz) amplitude modulations present in the envelope [4, 5]. This phase-locking was originally thought to reflect a strictly feed-forward process, but recent studies have found that it is also subject to top-down factors. For example, phase-locking is diminished when speech is unintelligible [6, 7], and is strengthened when the speaker's face is visible [8]. Additionally, in situations with multiple competing talkers, such as the classic 'cocktail party' task [9], the auditory system preferentially phase-locks to the envelope of the attended speech [10, 11], and tends to remain out of phase with the envelope of competing speech [12].

If these differences between the cortical responses to the envelopes of attended and unattended speech are visible in single-trial electroencephalographic (EEG) data, they could

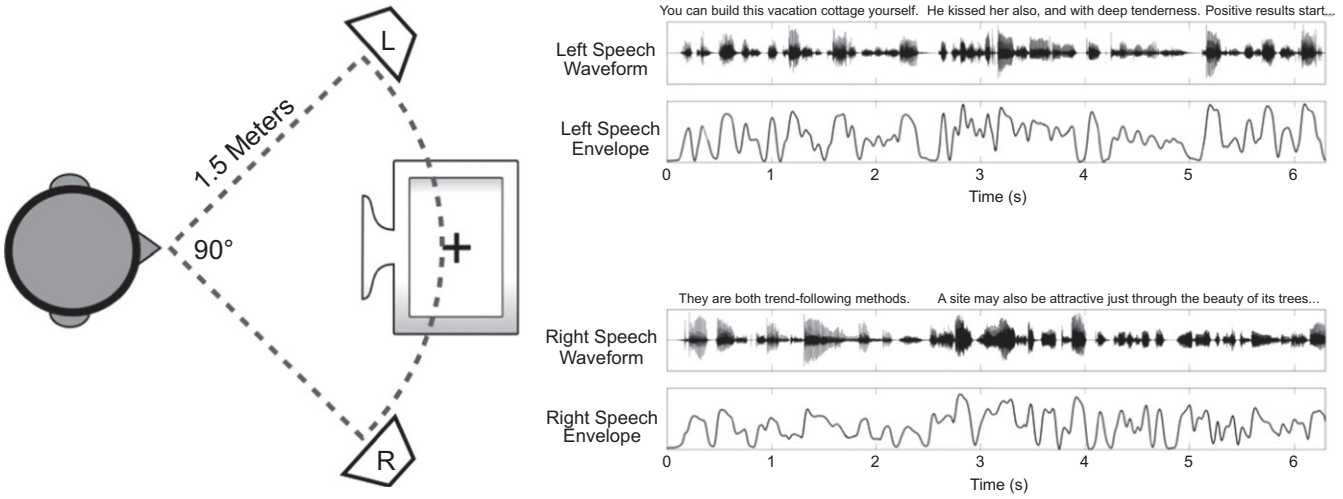

Figure 1. Layout and stimuli. Left: layout of the equipment during the task. Right: example of speech waveforms and envelopes from the first few seconds of a trial.

form the basis for a novel brain–computer interface (BCI). BCIs vary widely in their implementation, but the common goal is to use brain-generated signals to communicate with others or to control a computer interface [13]. Some BCIs allow users to signal intent by modulating some aspect of their brain activity, such as motor rhythms [14], but these often require considerable training. Other BCIs instead have subjects make choices by attending to one of several competing visual and/or auditory stimuli. By presenting each stimulus at a different time [15] or frequency [16, 17], evoked responses to each can be extracted and compared for signs of attention. While this method can allow for complex interfaces, such as a full BCI-controlled keyboard [15], it is constrained by the need for precisely-controlled artificial (e.g., flickered or modulated) stimuli. In contrast, an envelope-based attention BCI could operate on the amplitude modulations already present in complex naturalistic stimuli such as speech.

As a precursor to developing an envelope-based BCI, we determined in the present study whether the differences in neural responses to the envelopes of attended and unattended speech could be reliably observed in single-trial EEG data. Adult subjects performed a task in which they attended to one of two competing speakers while high-density EEG data were recorded. We extracted cortical responses to the envelopes of both speakers using cross-correlation, and then assessed our ability to decode each subject's side of attention as a function the duration of EEG data used to extract the responses. We found that envelope responses were sufficiently represented in EEG to decode side of attention from brief segments of data, with accuracy improving as data segment duration increased.

Furthermore, we wanted to compare classification performance using these envelope responses to classification performance using indicators of attention that have appeared in previous BCI studies. First, since the speakers were in different locations, we expected to see signs of attention in the EEG data's spectral content. The deployment of attention to the left or right side of space is associated with hemispheric lateralization of oscillatory power in several frequency bands [18–20]—particularly in the alpha band (8–12 Hz). In some studies, alpha power lateralization can be sufficiently robust to discriminate a subject's side of attention without further need to consider any stimulus-related brain activity [20, 21]. Second, some BCIs decode attention from changes in auditory steady-state responses (ASSRs) [17], as attention has been shown to boost ASSR magnitudes [22]. Thus, we amplitude modulated the left and right speech streams at 40 and 41 Hz in order to induce ASSRs in the EEG data. We found that classification accuracy using envelope responses greatly outperformed classification using either alpha lateralization or ASSR magnitude, further reinforcing the potential for envelope-based BCIs.

2. Methods

The data were also used in a previous study that examined how cortical entrainment to speech envelopes is involved in selectively attending to one of multiple speakers [12]. The current study shared some of its data pre-processing steps, but otherwise had distinct goals and analyses.

2.1. Participants

All experimental procedures were approved by the Institutional Review Board of the University of California, Irvine. Ten young adults (two female) between the ages of 21 and 29 volunteered to participate in the study, although one had to be excluded due to excessive EEG artifacts. All reported having normal hearing and no history of neurological disorder. Written informed consent was obtained from each subject prior to participation in the study.

2.2. Task and stimuli

Each participant sat in a sound-attenuated testing chamber facing a computer monitor flanked on either side by a

loudspeaker (figure 1). Before 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. During the trial, the left and right speakers played independent speech stimuli consisting of a series of spoken sentences taken from the TIMIT speech corpus [23]. To construct these speech stimuli, sentences were drawn from the corpus at random and concatenated until the total length of each channel exceeded 22 s, with silent gaps longer than 300 ms being reduced to 300 ms. No sentence was reused within experimental sessions. After constructing the stimuli, the left and right channels were sinusoidally amplitude-modulated at 40 and 41 Hz, respectively, in order to induce ASSRs. These modulation frequencies induce robust ASSRs [24, 25] and do not interfere with the intelligibility of the speech [26–28]. Envelopes for the speech were obtained by calculating the Hilbert transform of the stimuli, and then filtering the magnitude of the result with a passband of 2–30 Hz.

At the end of each trial, subjects were shown the transcript of a sentence they had heard during that trial. They were then required to indicate via a button press whether the sentence was played on the attended side. In practice, this task was very difficult unless subjects ignored the unattended side completely, as the memory load required to maintain both sides was prohibitive. Subjects were allowed to practice the task until their performance exceeded 80%, and were required to maintain that level throughout the experiment. Subjects completed 320 trials each (8 blocks, 40 trials per block), spread over 1–2 weeks, with the exception of one subject who only completed 240 trials due to equipment failure.

2.3. EEG recording and pre-processing

During the task, we recorded 128 channels of EEG using electrode caps, amplifiers, and software produced by Advanced Neuro Technology. Electrodes were placed following the international 10/5 system [29], and all channel impedances were kept below 10 kΩ. The EEG data were sampled at 1024 Hz with an online average reference. After the experiment, EEG data were exported into MATLAB (MathWorks, Natick, MA) for all further processing and analyses.

Each channel of EEG was filtered with a pass band of 1–50 Hz using a third order Butterworth filter. Filtering was conducted both forwards and then in reverse to eliminate phase shifts. The filtered data were then down-sampled to 256 Hz and segmented into individual trials which were 20 s long, beginning one second after the onset of the sentences. The delay between sentence onset and analysis window onset was necessary because neural onset responses are known to be large relative to envelope-related activity [1, 2]. Furthermore, since the left and right speech began simultaneously, they might have very briefly had correlated envelopes, which could impair later analyses. The segmented trials were visually inspected to exclude those with excessive artifacts (mean 16.6 trials per subject). The remaining data were then entered into the Infomax Independent Component Analysis algorithm available as part of the EEGLAB toolbox [30]. Components corresponding to artifacts such as eye movements and muscle activity were removed [31], and all remaining components were projected back into channel space for subsequent analyses.

2.4. EEG feature extraction

We extracted three different features of interest from the EEG data. First, we obtained envelope responses by calculating the cross-correlation functions [32] between each stimulus's envelope and the EEG channels. Cross-correlation measures the similarity of two time-series as a function of lag between them, and has been used previously to quantify neural responses to speech envelopes [1, 2, 12]. For two time-series $\boldsymbol{X}$ and $\boldsymbol{Y}$ of length N, the cross-correlation between $\boldsymbol{X}$ and $\boldsymbol{Y}$ at lag τ is defined as:

$$r(\boldsymbol{X}, \boldsymbol{Y}, \tau) = \frac{1}{N-\tau} \sum_{N-\tau}^{i=1} \frac{\left(x_i - \bar{\boldsymbol{X}}\right)\left(y_{i+\tau} - \bar{\boldsymbol{Y}}\right)}{\sigma_{\boldsymbol{X}}\sigma_{\boldsymbol{Y}}},$$

in which $\bar{\boldsymbol{X}}$ and $\bar{\boldsymbol{Y}}$ are the means and $\sigma_{\boldsymbol{X}}$ and $\sigma_{\boldsymbol{Y}}$ are the standard deviations of $\boldsymbol{X}$ and $\boldsymbol{Y}$. The second feature we extracted from each channel was a measure of power in the alpha band, calculated by performing the discrete Fourier transform on each trial and then summing power across all of the frequency bins between 8 and 12 Hz. The final feature we extracted from each channel was a measure of the ASSRs, calculated by again Fourier transforming each trial, but selecting just the frequency bins corresponding to the left (40 Hz) and right (41 Hz) modulators.

2.5. Within-subject classification

With the extracted EEG features, we attempted to build linear models that could predict to which speaker a subject attended on each trial. A graphical representation of this classification process appears in figure 2. First, all the trials for a given subject were randomly assigned to either a training set or a testing set, consisting of 75% and 25% of the total trials, respectively. Next, using just the training data, we attempted to find the most informative EEG channels for each feature. For the envelope cross-correlation feature, we computed the average cross-correlation functions for all attended stimuli in the training set, as well as the average cross-correlation functions for the unattended stimuli. The channels that were selected were the 15 which showed the greatest difference between their attended and unattended responses, as seen in figure 3. Using 15 channels was chosen as a balance between minimizing the complexity of the classifier while maximizing the information available to the classifier, as 15 channels could generally cover the peak responses. For the cross-correlation feature, we also performed the additional step of choosing the latencies at which the cross-correlation values should be entered into the classifier. We used the latencies corresponding to the three large peaks in the average attended minus unattended cross-correlation functions (see figure 3).

A similar process was used to find the most informative channels for the remaining two EEG features. For alpha

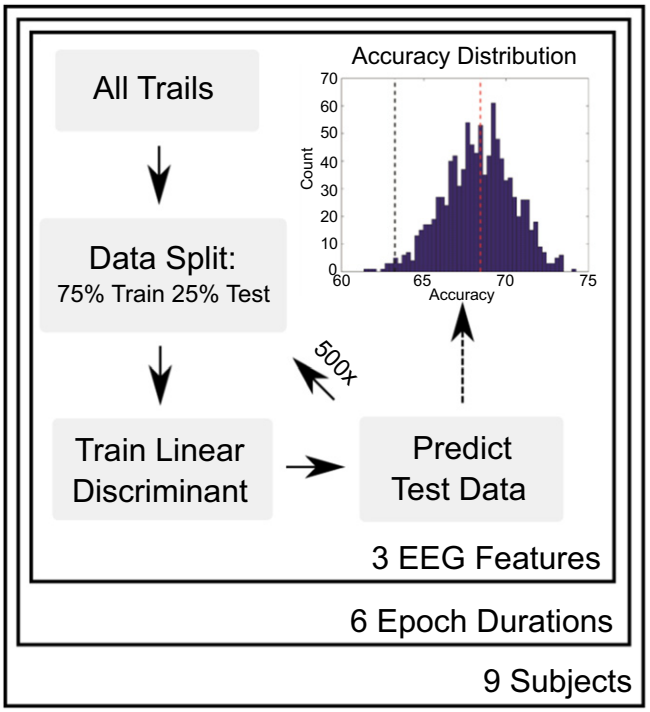

Figure 2. Within-subject classification process. A graphical representation of the within-subject classification process. For a given subject, epoch duration, and EEG feature (cross-correlations, alpha power, or ASSRs), epochs are randomized into training and test sets. A linear discriminant is formed using the training data, and is then used to predict the side of attention for each epoch in the test set. The process is repeated 500 times, with new random splits of training and test epochs. The accuracies of all iterations form a distribution (upper right). The mean of the distribution (red dashed line) is reported as the overall classification performance, and the accuracy is stated to be significantly above chance if the 5/6 percentile of the distribution (black dashed line) is above 50%.

power, we identified the most informative channels by subtracting the mean alpha power in each channel during 'attend left' trials in the training set from the mean of 'attend right' trials, and then selected those 15 channels where the differences were greatest (figure 4, left). For the ASSR feature, we used the 15 channels where the ASSR response magnitudes were greatest in the training data, averaged across the two modulation frequencies (figure 4, right).

Once we had determined which channels would be used for each feature, we proceeded to the classification step. Classification was performed using the linear discriminant classifier function built into MATLAB (classify.m). This classifier uses the training data to build an optimal hyperplane that separates the two conditions ('Attend left' and 'Attend right'). Once built, the hyperplane is used to predict the condition of each trial in the testing data. A separate classifier was built for each EEG feature of interest. For the cross-correlation feature, a total of 90 envelope-EEG cross-correlation values were entered into the classifier for each trial (15 channels, 3 latencies, 2 envelopes). For the alpha-power feature, the classifier received 16 values per trial: the alpha power at each of the 15 selected channels, as well as the difference of the average alpha power in the selected channels over the left hemisphere to those over the right hemisphere. This latter value was intended to better capture lateralization in alpha power in case the overall alpha power varied greatly across trials. Finally, for the ASSR feature the classifier received 30 values per trial (ASSR power at 15 channels, 2 frequencies).

The accuracy of each classifier was determined by comparing the classifier's prediction of the condition ('Attend left' *versus* 'Attend right') of each testing set trial to its true condition. That accuracy value (percent correct) comprised a single estimate of the classifier's performance. We then repeated the entire classification process starting from a new random split of training and test trials each time, until we had 500 such estimates, forming a distribution of classification accuracy values (see figure 2). Note that by using different training sets of data for each round of classification, different channels could potentially be chosen as the most informative for each iteration. However, in practice the most informative channels (and latencies for cross-correlations) were highly consistent across each iteration. From the distribution of 500 classification accuracy values, we used the mean to gauge the overall performance of the classifier. Classification accuracy was stated to be significantly above chance (one-way, $\alpha = .05$) if the 5th percentile of those accuracy estimates exceeded 50%.

In order to evaluate the relationship between the EEG epoch length and classification performance, we repeated the classification process for epoch lengths of 2, 3, 4, 5, 10, and 40 s. For epoch lengths shorter than 20 s (the length of trials in the original dataset), we divided each trial into multiple shorter epochs (i.e. one 20 s trial cut into five 4 s epochs). Each of the EEG measures (envelope-cross correlations, alpha power, and ASSRs) was then calculated on these new shorter epochs. For the epoch length of 40 s, we concatenated the EEG of two trials from the same condition. In order to account for multiple comparisons at the six different epoch lengths, we implemented a Bonferroni correction which changed the threshold for significance for any given epoch length to be set at the 5/6 percentile of the distribution instead of the 5th.

2.6. Cross-subject classification

While the classification process described above gives us an idea of how well a subject's direction of attention can be predicted after training a classifier to their own data, it would be beneficial to know how well their direction of attention could be predicted without tailoring the classifier to their individual data. Thus, we performed a second round of classification in which the classifiers were not trained on the individual subject's data. Instead, for each subject we used their entire data set for testing, and used all other subjects' data for training. Thus, the channels selected (and latencies for cross-correlations) would be those most informative across all other subjects.

In order to build a distribution of accuracy values similar to those of the individualized classifiers, we used a bootstrap sampling technique to determine which data epochs would be

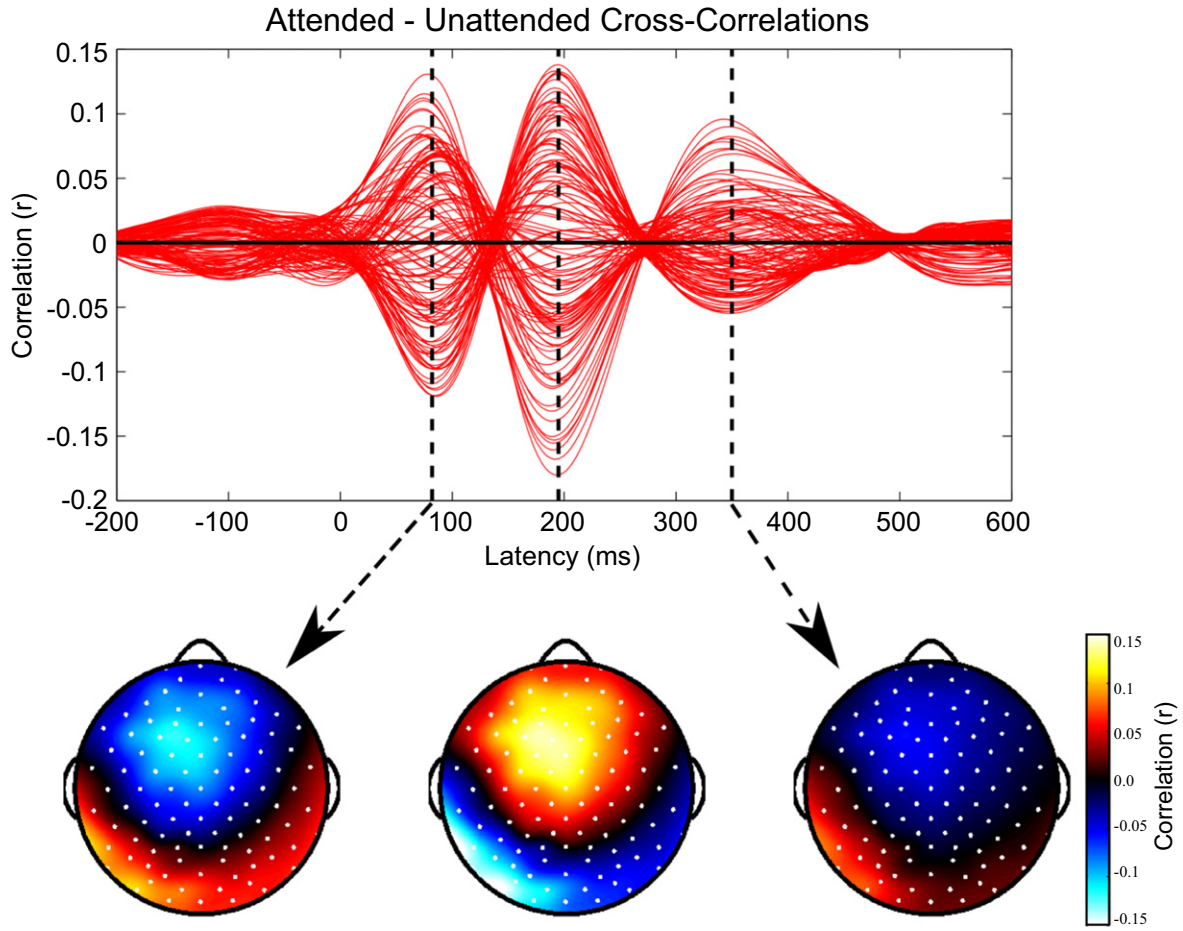

Figure 3. Cross-correlations: identifying channels and latencies of interest. For each subject, we calculate the average cross-correlation functions for the attended and unattended stimuli in the training data, and then plot their difference to identify channels and latencies where they are most distinct. Red lines indicate individual channels. Within the 15 channels with the largest magnitude differences between the attended and unattended cross-correlation functions, we find the three largest peaks in the difference function. Data shown for one representative subject.

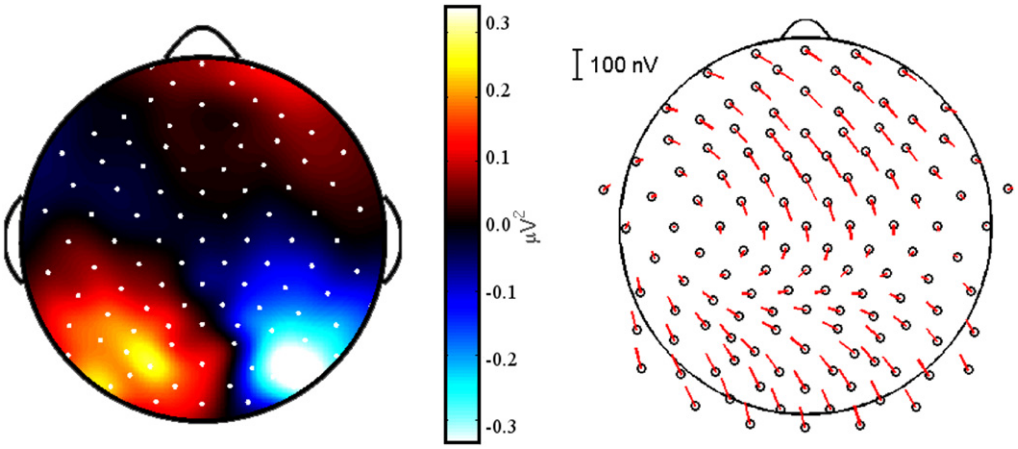

Figure 4. Alpha lateralization and ASSRs. Left: a topographic plot for a representative subject showing the difference in alpha (8–12 Hz) power between all 'Attend left' trials and 'Attend right' trials in the training set. A lateralization in alpha power is visible, with maximal differences measured in parietal electrodes. Right: the average ASSRs for a representative subject. Magnitude is indicated by the length of the line extending from each electrode, while the phase is indicated by the angle. ASSR topography was typical for EEG studies, with peaks in magnitude over frontal and occipital electrodes.

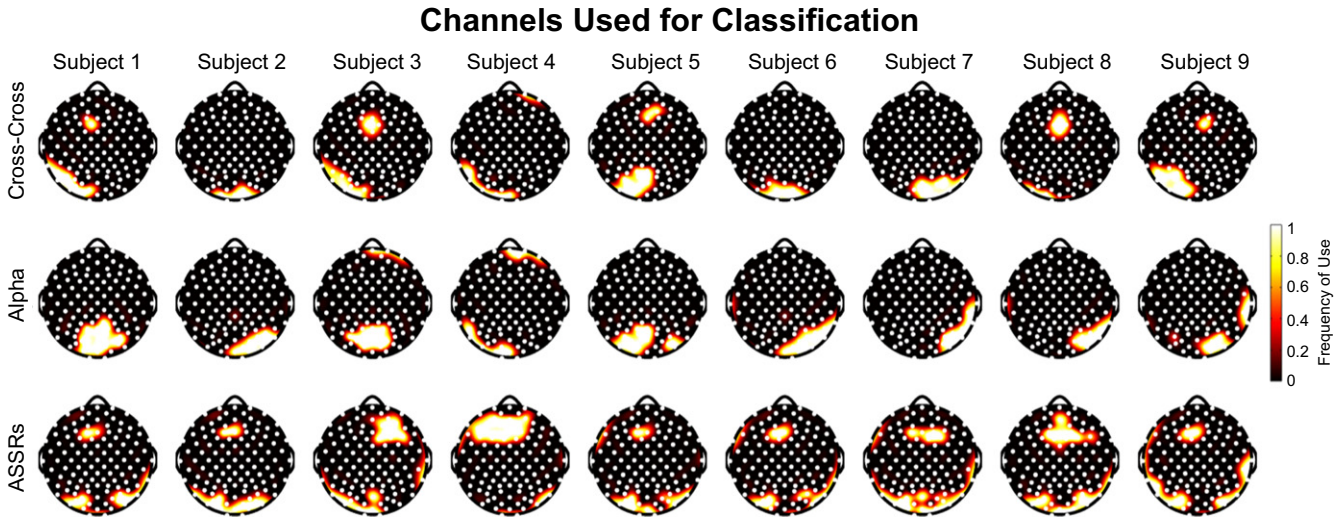

Figure 5. Selected channels. Topographies depicting the most informative channels for each subject and each feature. Frequency of use for each channel ranges from 1, meaning it was selected as one of the 15 most informative in every single train/test iteration, to 0, meaning that the channel never was selected as one of the most informative 15 channels.

used to train the classifier. This process entailed taking a sample of epochs with replacement from the training data of size *N*, where *N* was the total number of epochs in the training set. We then used that data to train a classifier that predicted the condition of epochs in the test subject's data. This process was repeated 500 times, with new samples of training data taken each iteration. As before, the accuracy values of the predictions formed a distribution which we treated similarly to the distributions described above.

3. Results

3.1. Behavior

Participants were able to exceed the required performance on the behavioral task throughout all experimental sessions (mean 82.45% correct). They reported the task as being challenging due to the effort required to maintain all of the novel sentences from the attended side in working memory.

3.2. EEG feature extraction

Using the training data for each subject, we calculated envelope cross-correlation functions that mirrored those observed in other studies [1–3, 12]. As seen in figure 5 (top), the channels where the attended and unattended cross-correlation functions were most distinct were located over frontal and temporal sites, which we have previously identified as consistent with sources in both early and later auditory areas [12]. Most subjects showed three distinct peaks in the difference function between the attended and unattended cross-correlations. The latencies of those peaks (around 90, 200, and 340 ms) corresponded to the latencies of well-known auditory evoked responses [33], and were highly similar across subjects.

Alpha power in each subject's training data showed the expected pattern of hemispheric lateralization differences between 'Attend left' and 'Attend right' trials, although those differences were generally weak. The channels selected for classification in each subject were typically located over parietal cortex, which is typical for auditory spatial attention tasks [20], but also included some neighboring occipital and temporal electrodes. Topographic plots showing which channels were selected for alpha power measures appear in figure 5 (middle).

Robust ASSRs were present in each subject's training data, but these ASSRs were not modulated by attention in any subject when comparing all 'Attend left' versus all 'Attend right' trials. The largest responses were recorded over frontal, occipital, and posterior temporal sites—consistent with previous ASSR studies [24, 34]. Topographic plots showing which channels were selected for ASSR measures appear in figure 5 (bottom).

3.3. Within-subject classification

We found that cross-correlation functions calculated from single epochs of EEG were highly effective in decoding which speaker had been attended, with classification accuracy exceeding chance for all subjects at all tested epoch lengths (figure 6 and table 1). At the shortest epoch length tested, 2 s, the average classification performance across subjects was 62.5%. That performance increased monotonically as epoch length increased, reaching 75% accuracy on average across subjects with 10 s of data. At 40 s, we saw evidence of ceiling effects on classifier performance, with three subjects above 95% classification accuracy. Classification accuracy differed greatly across subjects, with as much as a 20% difference in accuracy between the best- and worst-performing subjects. Those differences remained consistent across epoch lengths (i.e. the best- and worst-performing subjects were the same at all epoch durations).

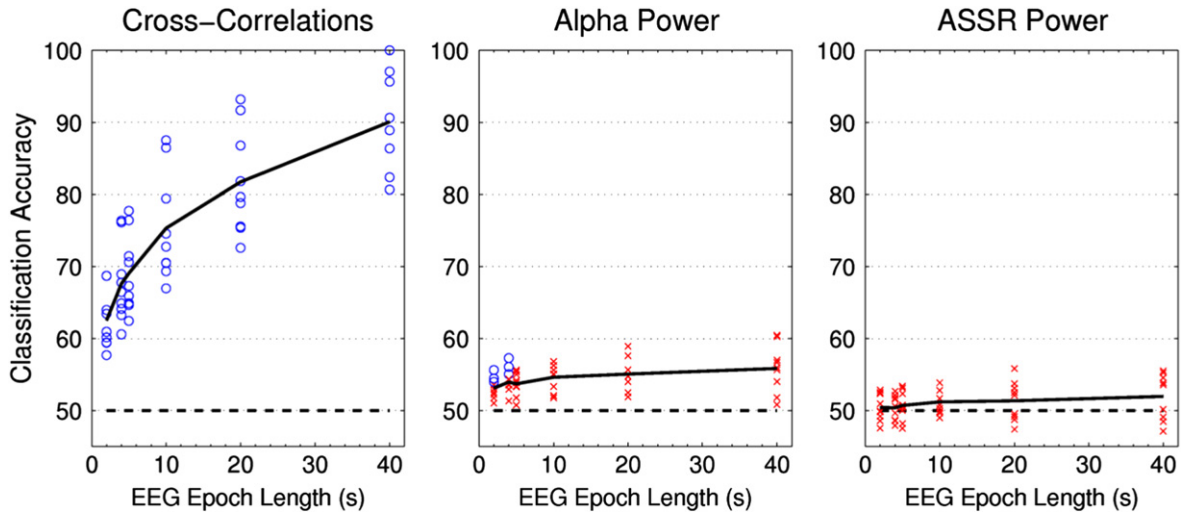

Figure 6. Within-subject classification results. Classification accuracy is plotted as a function of EEG sample length for each of the three features extracted from the EEG. Each point represents the mean classification accuracy for a single subject, while the solid black line indicates the average across all nine subjects. Chance is marked with the dashed line at 50% classification accuracy. Significantly above-chance accuracy values are marked by circles, while non-significant values are indicated by crosses.

In contrast to the classification based upon cross-correlations with the stimulus envelopes, we found that classification based on alpha power was poor. Mean classification performance never dipped below 50% for any subject or epoch length, indicating that there was some information about side of attention available in the alpha power. However, only a few subjects showed significantly above-chance classification accuracy, and those were still low and exclusively found at short epoch lengths. Classification based on ASSR magnitudes fared even worse, with accuracy never exceeding chance for any subject at any epoch length.

3.4. Cross-subject classification

When using other subjects' data for training, classification accuracies for the cross-correlation feature retained a similar shape to those of classifiers trained on the subject's own data, with accuracy increasing monotonically with epoch length (figure 7 and table 2). Overall classification performance was reduced as compared to the within-subject classifiers, ranging from 8% less at shorter epoch lengths to 10% less at longer epoch lengths, but still above chance for all subjects at all epoch lengths. When using the cross-subject data to train alpha power classifiers, the results closely mirrored those seen in within-subject classification. Alpha power could only significantly outperform chance at the shortest epoch lengths, and only in two subjects. As before, the ASSR classifiers failed to predict any subject's data better than chance at any epoch length.

4. Discussion

4.1. Cross-correlations

We were able to determine subjects' locus of attention using the cross-correlations between the speech envelopes and their EEG, with accuracy increasing as a function of epoch length. Encouragingly, classification performance for short epoch lengths was on par with or exceeded that seen in a comparable recent study using magnetoencephalography [35], a technology that is often used to measure neural speech responses but is impractical for most BCI purposes. Furthermore, we were also able to determine the subject's locus of attention from classifiers trained on other subjects' data, albeit with reduced performance, indicating that a BCI based on envelope cross-correlations would not necessarily have to be trained on a potential user's data prior to use. However, training the classifier on their own data would maximize performance.

The accuracy with which we were able to classify attention varied widely across subjects, with as much as a 20% difference in accuracy between the best and worst subjects at longer epoch lengths. Put another way, the classification accuracy using 2 s worth of EEG data from the best subject was equivalent in performance to 20 s of data from the worst subject. These differences between subjects are similar to those seen in other types of BCIs, where it has long been known that some subjects innately perform better with BCIs than others, and pre-training ability to use a BCI is a very strong predictor of post-training success [36]. However, in this task the individual differences would not be driven by a failure to learn how to modulate certain brain rhythms, but rather on differences in the robustness of the stimulus-related signals of interest. On *post-hoc* examination of the high and low performing subjects, the better performers had stronger cross-correlations between the speech envelopes and their EEG, and thus also would have had higher signal-to-noise ratios on individual epochs. Although we did not observe any notable differences between these groups in their behavioral performance, it would be interesting in future work to find out if their abilities diverged during more challenging multi-talker tasks.

Table 1. Within-subject classification results[1].

	2 s Epochs				4 s Epochs				`				10 s Epochs				20 s Epochs				40 s Epochs			
	CI−	Mean	CI+	SD	CI−	Mean	CI+	SD	CI−	Mean	CI+	SD	CI−	Mean	CI+	SD	CI−	Mean	CI+	SD	CI−	Mean	CI+	SD
Cross-correlations																								
S1	60.9	63.9	67.2	1.6	64.5	68.9	73.3	2.2	66.3	71.4	75.8	2.4	72.8	79.4	85.3	3.0	77.9	86.8	94.1	3.8	91.2	97.1	100.0	2.9
S2	64.8	68.7	72.0	1.8	71.4	76.3	80.5	2.4	72.5	77.7	83.4	2.7	81.3	86.5	92.7	3.2	85.4	91.7	97.9	3.2	91.3	95.7	100.0	3.0
S3	60.0	63.4	67.0	1.8	62.0	67.8	72.8	2.6	65.2	70.6	76.0	2.7	67.3	74.5	81.8	3.7	72.7	81.8	90.9	4.7	77.8	88.9	96.3	5.4
S4	55.2	59.5	63.3	2.0	57.8	63.2	68.6	2.8	58.4	64.6	70.2	3.2	61.4	69.3	77.3	4.3	64.4	75.6	86.7	5.5	72.7	86.4	95.5	6.6
S5	57.4	60.9	64.5	1.8	61.3	66.4	70.8	2.5	61.8	67.3	72.7	2.8	65.5	72.7	80.0	3.7	70.4	79.6	88.9	5.0	77.8	88.9	96.3	5.8
S6	64.9	68.7	72.5	2.0	70.7	76.1	81.1	2.6	70.8	76.4	82.0	2.8	81.8	87.5	93.2	3.1	86.4	93.2	97.7	3.2	95.5	99.5	100.0	2.3
S7	56.0	59.4	62.7	1.7	59.8	64.1	68.4	2.2	59.9	64.9	69.9	2.5	63.3	70.5	76.3	3.4	66.7	75.4	84.1	4.8	70.6	82.4	94.1	5.8
S8	54.0	57.7	61.2	1.8	55.6	60.6	66.3	2.6	56.8	62.4	67.4	2.7	60.5	66.9	74.2	3.6	62.9	72.6	82.3	5.0	67.7	80.6	90.3	6.4
S9	56.8	60.2	63.5	1.7	60.0	64.8	69.1	2.3	60.6	65.9	71.2	2.7	63.6	70.5	77.3	3.4	69.7	78.8	87.9	4.5	81.3	90.6	96.9	4.6
AVG	58.9	62.5	66.0	1.8	62.6	67.6	72.3	2.5	63.6	69.0	74.3	2.7	68.6	75.3	82.0	3.5	72.9	81.7	90.0	4.4	80.6	90.0	96.6	4.8
Alpha power																								
S1	49.7	52.9	56.1	1.6	50.3	55.0	59.2	2.3	48.9	54.2	59.4	2.7	48.8	56.3	63.0	3.6	45.0	54.0	64.5	5.3	42.0	57.0	69.0	7.1
S2	47.0	51.0	55.4	2.1	47.9	52.9	58.0	2.4	44.4	50.7	56.5	3.0	38.3	52.1	59.5	5.6	39.1	51.9	62.5	6.3	34.0	51.8	65.1	9.0
S3	52.1	55.6	59.1	1.8	52.1	57.3	62.5	2.5	49.4	55.4	60.9	2.9	47.3	55.6	64.0	4.2	43.6	55.6	66.8	5.7	41.6	56.7	71.8	8.2
S4	49.6	53.5	57.4	2.0	47.6	53.5	59.9	3.0	46.4	53.3	60.2	3.5	42.9	53.3	62.6	4.8	43.1	58.9	70.3	6.8	41.7	60.3	74.2	9.3
S5	50.4	54.0	59.0	2.1	48.8	54.4	59.2	2.7	48.7	55.6	61.7	3.2	47.3	55.6	64.0	4.3	43.4	54.8	68.0	6.2	41.6	60.4	75.6	8.2
S6	50.5	54.4	58.4	2.0	50.5	56.1	61.3	2.8	48.7	55.0	61.3	3.2	47.5	56.8	64.9	4.6	42.9	55.6	67.2	6.3	37.1	55.6	74.2	9.4
S7	49.6	52.5	55.8	1.6	48.9	53.6	57.7	2.2	47.9	53.4	58.5	2.6	47.0	55.0	61.6	3.7	44.3	54.7	65.0	5.3	33.0	54.0	66.0	8.0
S8	48.2	51.7	54.9	1.7	47.1	51.3	56.6	2.4	48.1	53.9	59.2	2.8	46.9	55.1	63.3	4.0	46.1	57.6	67.5	5.3	39.5	55.9	69.1	7.5
S9	49.2	52.2	55.3	1.5	47.0	51.3	56.3	2.3	46.2	51.8	56.8	2.6	43.3	51.8	58.7	3.7	40.1	52.5	61.7	5.3	34.9	50.8	66.8	7.9
AVG	49.6	53.1	56.8	1.8	48.9	53.9	59.0	2.5	47.6	53.7	59.4	2.9	45.5	54.6	62.4	4.3	43.1	55.1	65.9	5.8	38.4	55.8	70.2	8.3
ASSR power																								
S1	49.3	52.5	55.9	1.7	48.0	52.7	57.5	2.4	48.0	53.1	58.2	2.6	40.4	50.7	61.6	3.8	37.8	49.5	59.8	5.5	34.4	49.1	63.8	7.5
S2	48.4	52.3	56.5	2.1	46.3	52.1	57.9	3.0	45.6	52.3	58.0	3.2	38.2	49.9	61.6	4.2	40.6	53.1	65.6	6.3	38.1	55.5	70.7	9.0
S3	45.9	49.2	52.6	1.8	43.2	48.6	53.7	2.7	42.1	47.5	52.5	2.8	38.8	49.0	57.9	3.6	41.0	53.7	64.6	5.9	36.0	54.6	69.4	8.6
S4	49.0	52.8	56.7	1.9	45.6	50.6	56.0	2.8	47.2	53.4	59.6	3.1	40.1	52.8	64.7	4.5	38.8	52.1	65.4	6.5	35.4	53.5	71.7	9.4
S5	44.1	47.5	50.9	1.7	43.2	47.9	53.0	2.5	43.2	48.2	54.1	2.8	40.1	52.8	64.3	4.2	38.0	49.1	58.4	5.2	31.6	50.1	65.0	7.6
S6	44.3	48.6	52.4	2.1	42.5	48.6	54.0	2.9	43.8	50.0	56.7	3.2	36.9	49.6	62.3	4.6	35.1	48.7	60.1	6.5	30.3	48.5	66.6	9.0
S7	47.7	50.9	54.1	1.6	47.1	51.6	56.3	2.3	45.2	50.5	55.6	2.7	41.2	50.3	59.3	3.5	36.5	47.4	56.1	5.3	35.4	47.1	61.8	7.3
S8	46.6	49.8	53.0	1.7	45.2	49.6	54.1	2.4	44.8	50.4	56.0	2.8	40.4	51.7	63.0	4.1	42.9	55.8	65.5	5.8	37.7	53.8	66.7	7.8
S9	46.9	49.9	53.0	1.6	47.2	51.7	55.7	2.3	45.8	50.8	55.7	2.5	44.3	53.9	63.4	3.6	43.4	52.5	63.1	5.1	39.6	55.3	67.8	7.4
AVG	46.9	50.4	53.9	1.8	45.4	50.4	55.4	2.6	45.1	50.7	56.3	2.9	40.0	51.2	62.0	4.0	39.3	51.3	62.1	5.8	35.4	51.9	67.1	8.2

[1] Full classification results including the means and standard deviations of the accuracy distributions, as well as the lower and upper bounds of the 99% confidence intervals for the means.

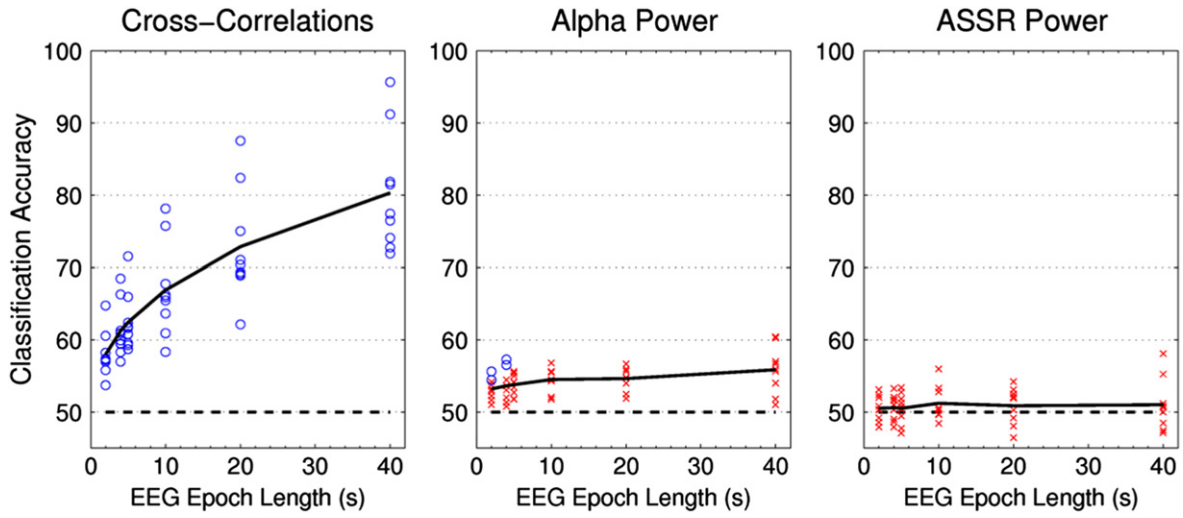

Figure 7. Cross-subject classification results. Classification accuracy is plotted as a function of EEG sample length for each of the three features extracted from the EEG. Each point represents the mean classification accuracy for a single subject, while the solid black line indicates the average across all nine subjects. Chance is marked with the dashed line at 50% classification accuracy. Significantly above-chance accuracy values are marked by circles, while non-significant values are indicated by crosses.

4.2. Alpha lateralization

Since alpha power showed the expected lateralization in the subject averages, it may seem puzzling at first why classification based on alpha power in single epochs was ineffective. The most likely explanation is that alpha lateralization is primarily associated with the *deployment* of spatial attention, not the *maintenance* of spatial attention. In this task, the most crucial time for deployment of spatial attention is at the very beginning of the trial, which is not included in our analysis window due to the problems that onset responses cause for the cross-correlation analyses. During our analysis window, the subjects are primarily maintaining attention at the cued location, which may not produce strong lateralization in alpha power. In fact, a similar cocktail party study found that alpha lateralization peaked 400–600 ms after sentence onsets, and was largely gone by 1000 ms (when our analysis window began) [37]. Subjects may have needed to briefly redeploy spatial attention at the transitions between sentences, which could explain why classification was able to exceed chance for a few subjects at short epoch lengths and why the alpha power was lateralized in the subject averages. However, this lateralization was clearly not robust enough in single epochs to produce useful classification of attention in this task.

Additionally, it is important to note that spatial location was not the only cue available for distinguishing between the two competing speech streams. Once the target speech stream had been segregated from the competitor, there are many other features besides spatial location that subjects can use to track the target speech stream, including pitch, timbre, and tempo. If the speaker's voice had been the same on both the left and the right, the spatial feature would likely have been much more salient to the subjects, and consequently may have produced much stronger lateralization of alpha power.

4.3. ASSRs

Attention did not affect the magnitudes of the ASSRs in the subject averages, and so it was unsurprising that the ASSR magnitudes did not help to classify attention in single epochs. ASSR insensitivity to attention was also reported in a recent similar study [10]. Since ASSRs have been shown to be sensitive to attention in the past [22, 38], and have been used to control a BCI [39], why were they not sensitive to attention here? We believe the difference lies in the fact that our stimuli were modulated speech utterances, whereas the studies in which ASSRs are affected by attention have used modulated tones, noise, or click trains. Modulated speech elicits much smaller ASSRs than modulated tones, noise, or reversed speech [25], suggesting that processing meaningful speech requires a suppression of the uninformative (and possibly interfering) amplitude modulation.

4.4. Conclusion

We have shown that neural responses to the envelopes of natural speech can be used to determine subjects' locus of attention, and thus could form the basis of a novel BCI. While the classification performance that we observed indicates that this BCI would not improve upon the information transfer rate in other BCIs, it would have the advantages of not requiring any training on the part of the subjects, and could be used with complex naturalistic stimuli such as speech. Additionally, while we only tested two-way classification, we could potentially increase the information transfer rate by increasing the number of speakers in the environment. In true cocktail-party scenarios, there may be dozens of competing speakers in the room, yet people are skilled at isolating the speaker of

Table 2. Cross-subject classification results[1].

	2 s Epochs				4 s Epochs				5 s Epochs				10 s Epochs				20 s Epochs				40 s Epochs			
	CI−	Mean	CI+	SD	CI−	Mean	CI+	SD	CI−	Mean	CI+	SD	CI−	Mean	CI+	SD	CI−	Mean	CI+	SD	CI−	Mean	CI+	SD
Cross-correlations																								
S1	57.6	60.6	63.8	1.6	61.9	66.3	70.7	2.2	61.2	65.9	71.1	2.5	69.5	75.7	80.9	3.2	73.5	82.4	89.7	4.1	82.4	91.2	100.0	4.2
S2	61.2	64.7	68.3	1.9	63.5	68.5	73.0	2.5	65.8	71.5	77.2	2.8	70.8	78.1	84.4	3.6	79.2	87.5	95.8	4.3	87.0	95.7	100.0	4.3
S3	52.0	55.8	59.6	1.9	54.7	59.4	64.9	2.7	53.8	59.3	65.2	2.9	52.7	60.9	69.1	4.1	58.2	69.1	80.0	5.6	59.3	74.1	88.9	7.6
S4	52.6	57.0	61.0	2.2	52.0	58.3	63.9	3.0	52.8	59.6	66.0	3.4	53.4	63.6	71.6	4.6	60.0	68.9	81.1	5.9	59.1	72.7	86.4	7.9
S5	53.6	57.3	60.7	1.8	54.4	59.5	64.6	2.7	56.4	61.8	67.5	2.9	59.1	65.5	73.6	3.8	61.1	70.4	81.5	5.3	66.7	81.5	92.6	6.7
S6	54.1	58.2	61.9	2.0	55.9	61.3	67.1	2.9	55.6	62.4	68.5	3.2	56.8	65.9	73.9	4.5	61.4	75.0	84.1	5.8	68.2	81.8	95.5	7.3
S7	54.2	57.4	60.5	1.7	56.6	60.9	65.2	2.2	56.3	61.6	67.0	2.6	59.0	66.2	72.7	3.5	60.9	71.0	79.0	4.7	64.7	76.5	91.2	6.3
S8	54.0	57.2	60.6	1.8	55.1	59.9	64.7	2.5	55.8	60.8	66.0	2.7	59.7	67.7	75.0	3.7	59.7	69.4	80.6	5.2	64.5	77.4	90.3	6.5
S9	50.3	53.7	57.1	1.8	52.1	57.0	61.2	2.4	53.4	58.7	63.6	2.7	50.8	58.3	65.2	3.6	53.0	62.1	72.7	5.3	57.8	71.9	84.4	7.2
AVG	54.4	58.0	61.5	1.8	56.2	61.2	66.2	2.5	56.8	62.4	68.0	2.9	59.1	66.9	74.0	3.9	63.0	72.9	82.7	5.1	67.7	80.3	92.1	6.4
Alpha power																								
S1	49.8	53.1	56.1	1.6	49.7	54.4	58.6	2.2	49.1	54.2	59.4	2.6	48.8	55.5	63.0	3.6	45.0	54.0	64.5	5.0	42.0	57.0	69.0	7.5
S2	46.8	51.0	55.2	2.1	45.3	50.8	56.3	2.8	44.9	51.8	58.4	3.2	42.5	52.1	60.6	4.6	41.3	51.9	66.8	6.4	29.6	51.8	69.5	9.8
S3	52.3	55.6	59.3	1.8	52.1	57.3	61.7	2.5	49.4	55.4	60.9	2.7	47.8	55.6	64.0	4.0	42.7	55.6	66.8	5.8	41.6	60.4	75.6	8.4
S4	49.4	53.8	57.9	2.1	46.7	53.1	59.0	3.0	45.8	52.7	60.2	3.4	44.0	54.5	62.6	4.6	43.1	56.7	68.0	6.5	37.1	60.3	74.2	9.7
S5	49.9	54.2	58.6	2.1	48.4	53.6	59.9	2.9	48.7	55.6	61.9	3.3	47.3	55.6	64.0	4.2	41.6	54.8	68.0	6.3	41.6	56.7	75.6	8.2
S6	50.5	54.4	58.1	1.9	50.8	56.5	62.5	2.8	48.1	55.0	60.7	3.3	47.5	56.8	66.1	4.5	44.0	55.6	69.5	6.6	41.7	55.6	74.2	9.0
S7	50.0	52.8	55.8	1.5	49.2	53.6	58.0	2.3	49.4	54.1	58.5	2.5	47.0	54.3	61.3	3.7	44.3	54.7	65.0	5.1	33.0	54.0	66.0	8.4
S8	48.1	51.7	54.9	1.7	47.4	52.0	57.0	2.4	48.1	53.4	59.6	2.8	46.1	54.3	62.5	4.0	44.4	55.9	65.8	5.2	39.5	55.9	69.1	7.7
S9	48.8	52.2	55.3	1.6	46.4	51.3	56.1	2.3	46.4	51.8	56.8	2.7	43.3	51.8	58.0	3.7	41.6	52.5	61.7	5.3	31.9	51.0	63.8	7.7
AVG	49.5	53.2	56.8	1.8	48.4	53.6	58.8	2.6	47.8	53.8	59.6	2.9	46.0	54.5	62.4	4.1	43.1	54.6	66.2	5.8	37.5	55.9	70.8	8.5
ASSR power																								
S1	49.9	52.9	56.3	1.7	48.1	53.1	57.5	2.4	47.3	52.7	57.9	2.6	38.7	49.0	59.3	3.8	37.0	48.1	58.4	5.6	32.4	47.1	64.8	7.9
S2	47.6	51.8	55.9	2.1	46.1	51.5	57.3	2.9	45.1	51.8	57.5	3.2	36.8	47.7	60.9	4.2	40.6	53.1	65.6	6.4	33.8	51.2	68.6	9.0
S3	45.3	49.0	52.1	1.8	43.0	48.4	53.5	2.7	42.1	47.1	52.5	2.8	39.4	49.6	59.7	3.8	41.0	51.9	63.7	5.8	36.0	50.9	65.7	8.3
S4	48.4	52.2	55.8	1.9	45.4	50.8	56.2	2.7	47.8	53.4	59.0	2.9	39.4	52.1	64.8	4.6	41.0	52.1	65.4	6.3	39.9	58.1	71.7	9.0
S5	44.1	47.7	51.3	1.8	42.8	47.7	52.8	2.5	42.7	47.7	53.2	2.8	39.4	49.6	61.0	4.0	38.0	49.1	60.3	5.4	32.6	47.4	62.3	7.4
S6	44.5	48.4	52.2	1.9	42.9	48.8	54.7	3.0	43.0	49.4	56.2	3.4	37.8	48.9	61.6	4.5	37.4	46.5	60.1	6.2	34.8	48.5	66.6	8.6
S7	47.6	50.6	53.9	1.6	46.9	51.4	56.0	2.2	46.1	51.3	57.0	2.8	42.5	52.6	64.7	4.0	38.7	50.3	60.4	5.2	32.4	50.0	64.7	7.8
S8	46.6	50.1	53.8	1.8	45.3	50.1	54.6	2.4	44.8	50.4	55.6	2.7	39.7	49.8	62.3	4.1	42.9	54.2	65.5	5.7	37.7	50.6	66.7	7.9
S9	47.0	50.2	53.1	1.6	47.3	51.8	56.1	2.2	46.2	50.8	55.1	2.4	44.7	55.3	64.8	3.4	43.4	52.5	61.6	5.0	39.6	55.3	67.8	7.1
AVG	46.8	50.3	53.8	1.8	45.3	50.4	55.4	2.6	45.0	50.5	56.0	2.8	39.8	50.5	62.1	4.1	40.0	50.9	62.3	5.8	35.5	51.0	66.5	8.1

[1] Full classification results including the means and standard deviations of the accuracy distributions, as well as the lower and upper bounds of the 99% confidence intervals for the means.

their choice. If people can do this behaviorally, there is good reason to believe that we could similarly isolate the neural response to the attended speaker, too. Thus, the upper limit of the information transfer rate for this BCI may be determined by the number of uncorrelated speech stimuli that can be presented at once.

Acknowledgements

We thank Robert Coleman for his suggestions regarding data classification. This work was supported by grants from the Army Research Office (ARO 54228-LS-MUR) and the National Institutes of Health (2R01-MH68004).

References

[1] Aiken S J and Picton T W 2008 Human cortical responses to the speech envelope *Ear Hear.* **29** 139–57
[2] Abrams D A, Nicol T, Zecker S and Kraus N 2008 Right-hemisphere auditory cortex is dominant for coding syllable patterns in speech *J. Neurosci.* **28** 3958–65
[3] Lalor E C and Foxe J J 2010 Neural responses to uninterrupted natural speech can be extracted with precise temporal resolution *Eur. J. Neurosci.* **31** 189–93
[4] Luo H and Poeppel D 2007 Phase patterns of neuronal responses reliably discriminate speech in human auditory cortex *Neuron* **54** 1001–10
[5] Giraud A-L and Poeppel D 2012 Cortical oscillations and speech processing: emerging computational principles and operations *Nat. Neurosci.* **15** 511–7
[6] Ahissar E, Nagarajan S S, Ahissar M, Protopapas A, Mahncke H and Merzenich M M 2001 Speech comprehension is correlated with temporal response patterns recorded from auditory cortex *Proc. Natl Acad. Sci. USA* **98** 13367–72
[7] Peelle J E, Gross J and Davis M H 2012 Phase-locked responses to speech in human auditory cortex are enhanced during comprehension *Cereb. Cortex* **23** 1378–87
[8] Zion Golumbic E, Cogan G B, Schroeder C E and Poeppel D 2013 Visual input enhances selective speech envelope tracking in auditory cortex at a 'cocktail party' *J. Neurosci.* **33** 1417–26
[9] Cherry E C 1953 Some experiments on the recognition of speech, with one and with two ears *J. Acoust. Soc. Am.* **25** 975–9
[10] Ding N and Simon J Z 2012 Neural coding of continuous speech in auditory cortex during monaural and dichotic listening *J. Neurophysiol.* **107** 78–89
[11] Zion Golumbic E M *et al* 2013 Mechanisms underlying selective neuronal tracking of attended speech at a 'cocktail party' *Neuron* **77** 980–91
[12] Horton C, D'Zmura M and Srinivasan R 2013 Suppression of competing speech through entrainment of cortical oscillations *J. Neurophysiol.* **109** 3082–93
[13] Birbaumer N and Cohen L G 2007 Brain-computer interfaces: communication and restoration of movement in paralysis *J. Physiol.* **579** 621–36
[14] Wolpaw J R and Mcfarland D J 1994 Multichannel EEG-based brain-computer communication *Electroencephalogr. Clin. Neurophysiol.* **90** 444–9
[15] Farwell L A and Donchin E 1988 Talking off the top of your head: toward a mental prosthesis utilizing event-related brain potentials *Electroencephalogr. Clin. Neurophysiol.* **70** 510–23
[16] Kim D-W, Cho J-H, Hwang H-J, Lim J-H and Im C-H 2011 A vision-free brain-computer interface (BCI) paradigm based on auditory selective attention *Conf. Proc. IEEE Eng. Med. Biol. Soc.* **2011** 3684–7
[17] Desain P, Hupse A M G, Kallenberg M G J, Kruif B J D and Schaefer R S 2006 Brain-computer interfacing using selective attention and frequency-tagged stimuli *Proc. 3rd Int. BrainComputer Interface Work. Train. Course* pp 2–3
[18] Worden M S, Foxe J J, Wang N and Simpson G V 2000 Anticipatory biasing of visuospatial attention indexed by retinotopically specific alpha-band electroencephalography increases over occipital cortex *J. Neurosci.* **20** RC63 pp 1–6
[19] Thut G, Nietzel A, Brandt S A and Pascual-Leone A 2006 Alpha-band electroencephalographic activity over occipital cortex indexes visuospatial attention bias and predicts visual target detection *J. Neurosci.* **26** 9494–502
[20] Thorpe S, D'Zmura M and Srinivasan R 2012 Lateralization of frequency-specific networks for covert spatial attention to auditory stimuli *Brain Topogr.* **25** 39–54
[21] Srinivasan R, Thorpe S, Deng S, Lappas T and Zmura M D 2009 Decoding attentional orientation from EEG spectra *Lecture Notes in Computer Science* pp 176–83
[22] Ross B, Picton T W, Herdman A T and Pantev C 2004 The effect of attention on the auditory steady-state response *Neurol. Clin. Neurophysiol.* **22** 1–4
[23] Garofolo J, Lamel L, Fisher W, Fiscus J, Pallett D, Dahlgren N and Zue V 1993 *TIMIT Acoustic-Phonetic Continuous Speech Corpus* (Philadelphia, PA: Linguistic Data Consortium)
[24] Picton T W, John M S, Dimitrijevic A and Purcell D W 2003 Human auditory steady-state responses *Int. J. Audiol.* **42** 177–219
[25] Deng S and Srinivasan R 2010 Semantic and acoustic analysis of speech by functional networks with distinct time scales *Brain Res.* **1346** 132–44
[26] Miller G A and Licklider J C R 1950 The intelligibility of interrupted speech *J. Acoust. Soc. Am.* **22** 167–73
[27] Drullman R, Festen J M and Plomp R 1994 Effect of reducing slow temporal modulations on speech reception *J. Acoust. Soc. Am.* **95** 2670–80
[28] Drullman R, Festen J M and Plomp R 1994 Effect of temporal envelope smearing on speech reception *J. Acoust. Soc. Am.* **95** 1053–64
[29] Oostenveld R and Praamstra P 2001 The five percent electrode system for high-resolution EEG and ERP measurements *Clin. Neurophysiol.* **112** 713–9
[30] Delorme A and Makeig S 2004 EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis *J. Neurosci. Methods* **134** 9–21
[31] Jung T-P, Makeig S, Humphries C, Lee T-W, McKeown M J, Iragui V and Sejnowski T J 2000 Removing electroencephalographic artifacts by blind source separation *Psychophysiology* **37** 163–78
[32] Bendat J S and Piersol A G 1986 *Random Data: Analysis and Measurement Procedures* vol 3 (New York: Wiley)
[33] Picton T W, Hillyard S A, Krausz H I and Galambos R 1974 Human auditory evoked potentials. I. evaluation of components *Electroencephalogr. Clin. Neurophysiol.* **36** 179–90
[34] Herdman A T, Lins O, Van Roon P, Stapells D R, Scherg M and Picton T W 2002 Intracerebral sources of human auditory steady-state responses *Brain Topogr.* **15** 69–86
[35] Koskinen M, Viinikanoja J, Kurimo M, Klami A, Kaski S and Hari R 2012 Identifying fragments of natural speech from

the listener's MEG signals *Hum. Brain Mapp.* **1489** 1477–89

[36] Neumann N and Birbaumer N 2003 Predictors of successful self control during brain-computer communication *J. Neurol. Neurosurg. Psychiatry* **74** 1117–21

[37] Kerlin J R, Shahin A J and Miller L M 2010 Attentional gain control of ongoing cortical speech representations in a 'cocktail party' *J. Neurosci.* **30** 620–8

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

[39] Kim D-W, Hwang H-J, Lim J-H, Lee Y-H, Jung K-Y and Im C-H 2011 Classification of selective attention to auditory stimuli: toward vision-free brain-computer interfacing *J. Neurosci. Methods* **197** 180–5